An Improved Two-Stage Procedure to Compare Hazard Curves

Zhongxue Chen[1]*, Hanwen Huang[2]*, and Peihua Qiu[3]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington. 1025 E. 7th street, Bloomington, IN, 47405, USA.

[2]Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia. Athens, GA 30602, USA.

[3]Department of Biostatistics, College of Public Health & Health Professions and College of Medicine, University of Florida. Gainesville, FL 32611, USA.

*Authors contribute equally

Running title: Improved two-stage procedure for survival data

Abstract

Qiu and Sheng has proposed a powerful and robust two-stage procedure to compare two hazard rate functions. In this paper we improve their method by using the Fisher test to combine the asymptotically independent p-values obtained from the two stages of their procedure. In addition, we extend the procedure to situations with multiple hazard rate functions. Our comprehensive simulation study shows that the proposed method has a good performance in terms of controlling the type I error rate and of detecting power. Two real data applications are considered for illustrating the use of the new method.

Key words: Bonferroni correction; crossing; Fisher test; hazard rate; survival data

1. Introduction

In survival data analysis, one common task is to compare two or more hazard curves. To this end, several nonparametric methods, including the log-rank (LR), Gehan–Wilcoxon and Peto–Peto tests, among several others, have been proposed in the literature (Lee and Wang, 2003). However, the performance of these methods depends on the true alternatives. For instance, the LR test is more powerful than other methods if the hazard functions under investigation are parallel. But, when such a proportional hazards assumption is violated, the LR test may lose power dramatically. To circumvent this limitation, some methods that are robust to the violation of the proportional hazards assumption have been proposed in the literature (Cheng et al., 2009; Li et al., 2015; Lin and Wang, 2004; Liu et al., 2007; Mantel and Stablein, 1988; Moreau et al., 1992; Park and Qiu, 2014; Qiu and Sheng, 2008).

The method proposed by Qiu and Sheng (called the QS test hereafter) is a two-stage procedure designed for comparing two hazard rate functions (Qiu and Sheng, 2008). In the first stage, the LR test is performed with the hope that two parallel hazards can be distinguished in the case when they are not identical. In the second stage, a test statistic (denoted as NP) is constructed which has the following properties: it is asymptotically independent of the one used in the first stage, and it is powerful when the two hazard functions cross each other. Let α be the overall significance level; α_1 and α_2 the significance levels for stages one and two, respectively. Denote p_1 and p_2 the p-values from the first and the second stages, respectively. The QS test performs as follows. In stage one, if $p_1 \leq \alpha_1$, then we reject the null hypothesis that the two hazard functions are the same and stop the testing procedure; otherwise, go to stage two. In stage two, if $p_2 \leq \alpha_2$, then we reject the null hypothesis mentioned above;

otherwise, the QS test fails to reject the null hypothesis. Therefore, asymptotically we have the following relationship among the significance levels:

$$\alpha_1 + \alpha_2(1-\alpha_1) = \alpha. \quad (1)$$

There are infinitely many different choices for α_1 and α_2 for a given overall significance level α in (1). In the original QS test, for convenience, α_1 and α_2 are chosen to be

$$\alpha_1 = \alpha_2 = 1 - \sqrt{1-\alpha}\,. \quad (2)$$

The overall p-value for the QS test can be defined by

$$p - value = \begin{cases} p_1, & p_1 \le \alpha_1 \\ \alpha_1 + p_2(1-\alpha_1), & otherwise \end{cases}. \quad (3)$$

There are some interesting characteristics associated with the QS test. First, the overall p-value and power depend on the choices of α_1 and α_2, and they also depend on the preset significance level α. Second, from (3), it is obvious that the overall $p - value \ge \min(p_1, \alpha_1)$. This may be a shortcoming if one views the p-value as the degree of evidence against the null hypothesis (e.g., a smaller p-value implies a stronger evidence). Third, the p-value and power of this test depend on the order of the two stages. Fourth, the QS test was designed for situations with two hazard curves only. In practice, one may have multiple hazard curves to compare; therefore, it is desirable to extend the QS test to cases with multiple hazard curves.

To overcome the aforementioned shortcomings of the QS test, and more importantly, to increase its detecting power and make it suitable for handling multiple hazard curves, we propose an improvement of the QS test. Through a comprehensive simulation study, we will show that the new method is more powerful than the QS test under many different situations. We will also use three real data examples to illustrate the use of the proposed method.

2. Method

2.1. The QS test

Suppose we have two treatment groups with n_j (j=1,2) subjects and the total number of subjects is $n=n_1+n_2$. We denote the D distinct ordered failure times as t_1, t_2, …, t_D. At each failure time t_i (i=1,2,…, D), d_{ij} and Y_{ij} are the number of events and the number of subjects at risk for group j. Let $d_i=d_{i1}+d_{i2}$, and $Y_i=Y_{i1}+Y_{i2}$. The LR test statistic used in the first stage of the QS test can be written as

$$U = \sum_{i=1}^{D} w_{i1}(d_{i1} - Y_{i1}\frac{d_i}{Y_i}) \Big/ \sqrt{\sum_{i=1}^{D} \frac{Y_{i1}}{Y_i}\frac{Y_{i2}}{Y_i}\frac{Y_i-d_i}{Y_i-1} d_i}, \quad (4)$$

where $w_{i1} = 1$ here.

To construct the test statistic in the second stage, we need the following notations. For each j (j=1,2), let T_{kj} (k=1,2,…,n_j) denote the event time of the kth subject in group j which has the cumulative distribution function F_j, and C_{kj} be the censoring time which has the cumulative distribution function G_j. Denote the survival functions for the survival time and censoring time as $S_j(s)= 1-F_j(s)$ and $L_j(s)= 1-G_j(s)$, respectively. Let $X_{kj} =\min(T_{kj}, C_{kj})$ be the observed survival or censored time, $\delta_{kj} = I(T_{kj} < C_{kj})$ be the censoring indicator, and $\pi_j(s) = P(X_{kj} > s) = S_j(s)L_j(s)$ be the survival function of the observed time. Then, under the null hypothesis that the two hazard functions are the same, we have $F_1 = F_2$ (or equivalently, $S_1 = S_2$).

The test statistic used in the second stage of the QS test is a weighted log-rank test defined by

$$V = sup_{D_\varepsilon \le m \le D-D_\varepsilon}(V_m), \quad (5)$$

where $m = [Dr]$ denotes the integer part of Dr, for any $r \in [\varepsilon, 1-\varepsilon]$, $0 < \varepsilon < 0.5$ is a given small number, $V_m = \sum_{i=1}^{D} \widehat{w}_{i2}^{(m)}(d_{i1} - Y_{i1}\frac{d_i}{Y_i}) \Big/ \sqrt{\sum_{i=1}^{D} \widehat{w}_{i2}^{(m)} \frac{Y_{i1}}{Y_i}\frac{Y_{i2}}{Y_i}\frac{Y_i-d_i}{Y_i-1} d_i}$,

$$\widehat{w}_{i2}^{(m)} = \begin{cases} -1, & if\ i = 1,2,\dots,m \\ \hat{c}_m, & otherwise \end{cases},$$

$\hat{c}_m = \sum_{i=1}^{m} \frac{\hat{L}_1(t_i)\hat{L}_2(t_i)}{\left(\frac{n_1}{n}\right)\hat{L}_1(t_i)+\left(\frac{n_2}{n}\right)\hat{L}_2(t_i)} \Delta\hat{S}(t_i) \,/\, \sum_{i=m+1}^{D} \frac{\hat{L}_1(t_i)\hat{L}_2(t_i)}{\left(\frac{n_1}{n}\right)\hat{L}_1(t_i)+\left(\frac{n_2}{n}\right)\hat{L}_2(t_i)} \Delta\hat{S}(t_i)$, and $\hat{L}_1$, $\hat{L}_2$, and $\hat{S}$ are the Kaplan-Meier estimates (Kaplan and Meier, 1958) of L_1, L_2 and S, respectively. The p-value for V can be approximated by a bootstrap procedure.

For the two statistics U and V used in the two stages of the QS test, we have the following nice property (Qiu and Sheng, 2008).

Theorem 1. The two statistics U and V in (4) and (5) are asymptotically independent under the null hypothesis that the two hazard functions in question are the same.

2.2. The proposed test

Since the two statistics U and V are asymptotically independent, the two p-values obtained by the two statistics are also asymptotically independent and identically distributed from 0 and 1 under the null hypothesis that the two hazard curves are equal. Based on this fact, we propose to obtain an overall p-value using the Fisher-test method since it is more robust in the sense that it has reasonable power when either one of the two p-values is small (Chen, 2011; Chen and Nadarajah, 2014; Fisher, 1932; Owen, 2009). By this method, the overall p-value is defined by

$$p_F = H^{-1}(-2\ln[\,p_1 p_2\,]), \qquad (6)$$

where H is the survival function of a random variable that has a chi-square distribution with degrees of freedom 4 based on the Fisher's theorem (Fisher, 1932).

Method (6) can be extended to cases with K ($K \geq 2$) hazard curves easily, in which we can make pairwise comparisons for a total of $K(K-1)/2$ pairs of curves. For each of the

comparisons, we can calculate its overall p-value using (6). Then, the Bonferroni procedure for multiple comparisons can be used to determine whether the null hypothesis should be rejected or not. More specifically, if the smallest p-value from the $K(K-1)/2$ comparisons is less than the given significance level divided by $K(K-1)/2$, then the null hypothesis is rejected; otherwise it is not rejected. It should be pointed out that here we don't use a global test for the overall comparison although the adoption of Bonferroni correction guarantees the controlling of family-wise error rate. It is possible that some global tests, if exist, may reject the null hypothesis at a given significance level but all of the adjusted p-values from our method are greater than that nominal level.

There are some good properties for the proposed test. Figure 1 shows the acceptance regions of the original QS test and the proposed method based on the Fisher test, both with a significance level of 0.05 in cases when K=2. It clearly shows that the QS test would reject the null hypothesis when either p_1 or p_2 is less than 0.0253. In contrast, the proposed method based on the Fisher test can reject the null hypothesis in certain cases when both p_1 and p_2 are larger than 0.0253 (i.e., (p_1, p_2) in the shaded region in Figure 1). In other words, for certain alternative hypotheses that both U and V have considerable powers to detect, the proposed method based on the Fisher test would be more powerful than the original QS test. In practice, such alternative hypotheses correspond to cases when the two hazard curves across at an early or late time, which are common in applications. On the other hand, for the alternative hypotheses that one of U and V has little power to detect but the other one has a good power, the QS test may perform slightly better. These alternative hypotheses usually correspond to cases when the two hazard curves cross at a quite middle time point or they are quite parallel

to each other, which are less common in practice. In cases when there are multiple hazard curves, it will be shown that the proposed method is much more powerful than the QS test in most scenarios.

Results

3.1. Simulation study

To assess the performance of the proposed test, we conduct a comprehensive simulation study to compare it with the QS tests. The distributions of the survival time (T) are assumed to be uniform ($U: U(\theta - a, \theta + a)$), exponential ($E: \exp(a) + \theta$), or log-normal (LN) ($LN: LN(\theta, a)$); the corresponding censoring time (C) follows a uniform ($U(\theta - a, \theta + a + 2(1-2p)/p)$), exponential ($E: \exp(a\frac{p}{1-p}) + \theta$), and log-uniform (LU) ($LU: LU(\theta + aU(-2, -2 + 2/p))$), respectively. Here p is the expected censoring rate. In the simulation we consider different values for p: 0, 0.2, 0.4, and 0.6. When estimating the empirical type I error rate, we assume the distributions for the survival times are identical for all groups. When assess the power, in addition to cases when all groups have the same types of distribution, we also consider cases when treatment groups have different types of distributions. More specifically, we also consider the following cases: some groups have uniform distributions and the others have exponential distributions (denoted as U+E), some groups have uniform distributions and the others have log-normal distributions (denoted as U+LN), and some groups have exponential distributions and the others have log-normal (denoted as E+LN). We compare the proposed method (denoted as New) with the QS two-stage test using default values $\alpha_1 = \alpha_2 = 1 - \sqrt{1-\alpha}$ (denoted as TS), the log-rank test (denoted as LR), and the test used in the second stage

of the QS test (denoted as NP). To see how the significance levels α_1, α_2 in the two stage affect the performance of the QS approach, we also consider two QS tests with different values for $\alpha_1$and α_2. Specifically, we chose $\alpha_1 = 0.01$, $\alpha_2 = 0.0404$ (denoted as TS1), and $\alpha_1 = 0.0404$, $\alpha_2 = 0.01$ (denoted as TS2).

We choose sample sizes to be 50 for both groups. We use the R package TSHRC with default setting (i.e, $\varepsilon = 0.1$, and the number of bootstrap samples is 1000) to get the p-values for TS, LR, and NP. To compute the empirical type I error rate and the power, we use a significance level of 0.05 and compute the proportion of rejections out of 1000 replicated simulations.

First, we consider the cases when there are only two hazard curves. Table 1 below reports the empirical type I error rates for each method. It can be seen that all methods considered can control the type I error rate well in different cases, although the methods TS, NP, and New tend to have slightly smaller empirical type I error rates in certain cases when the censoring rates are low.

Table 2 gives the empirical power values for each method under different settings. It shows that (i) in cases when both LR and NP have some power to detect the difference, the proposed test (i.e., New) is usually more powerful than the original QS test (i.e., TS), and (ii) in cases when only one of LR and NP has power, the QS test is slightly better than the proposed test. These results are consistent with our expectations demonstrated in Figure 1.

Next, we consider situations when we have four hazard curves. Similar to the proposed test, for comparison of multiple hazard curves, the p-values from the pairwise QS tests can be used to determine whether we should reject the overall null hypothesis or not. In other words, if the smallest p-value of the QS test obtained from the $K(K-1)/2$ pairwise comparisons is less

than the given significance level divided by $K(K-1)/2$, then the over null hypothesis that all hazard curves are the same is rejected; otherwise the null is not rejected. It should be pointed that the LR test can be applied directly to cases when K>2. The method NP is not included here since it requires an enormous number of bootstrap samples to accurately estimate its p-value, which is not feasible in this example with relatively small sample sizes.

Table 3 reports the empirical sizes of the method based on the QS tests (TS, TS1, and TS2), the LR test and the proposed method New. Again, all of the methods can control the type I error rate reasonably well.

Table 4 reports the empirical power values for each method under different settings. It can be seen that in many situations, the proposed test has larger power values than the LR test and the QS tests. Sometimes, the differences of the power values between the proposed test and the QS tests are big. For example, when the survival times are all from the uniform distributions for the four hazard curves, regardless of the censoring rates, the proposed method can detect the difference almost all the time. In contrast, the LR and QS methods have much lower powers, especially when the censoring rates are high (e.g., p=0.4 and 0.6).

Among those TS tests with different values for α_1 and α_2, in general, as expected, under situations where LR test is more powerful than the NP test, TS2 is more powerful than TS1; on the other hand, when LR is less powerful than the NP test, TS1 is more powerful than TS2. If there are two groups to be compared (Tables 2 and S2), we observed that under some conditions, the TS1 or TS2 may have larger empirical power values than the proposed test. However, in general, the new test has reasonable power under all conditions considered in the simulation: it has similar or larger power values compared with the maximum power values

from TS, TS1 and TS2. When there are four hazard rate functions to be compared (Tables 4 and S4), the new test usually has larger power values than those from TS, TS1, and TS2. In addition, with different values of α_1 and α_2, the empirical power value of the TS tests (e.g., TS1 vs. TS2) may change dramatically.

To see how sample sizes affect the performance of the new test and the TS tests, we also simulated data with sample size 100 per group. We keep all of the other parameters the same but use significance level $\alpha = 0.01$. The simulation results were reported in the Supplementary Tables S1-S4. We have similar observations as those from Tables 1-4: in general, the proposed test is preferred as it has reasonable power under all of the situations considered.

3.2. Real data applications

To illustrate the use of the proposed test, we applied it to three data sets. The first data set was obtained from a study about the tumorigenesis of a drug (Mantel et al., 1977) which has been analyzed by Qiu and Sheng (Qiu and Sheng, 2008). In this study, rats were taken from 50 distinct litters. One rat from each litter was randomly selected and given the drug, and another two rats were selected as controls and were given a placebo. There were 29 and 81 censored observations in the treatment group and the control group, respectively. The two p-values in the first and the second stages of the QS test were 0.0034 and 0.051, respectively. If we use the commonly used significance level 0.05, the overall p-value from the QS test will be 0.0034 and the null hypothesis is then rejected. However, if we choose the significance level to be 0.005, then the overall p-value will be 0.053, which is larger than the preset significance level. Therefore, we will not reject the null hypothesis in such cases. As a comparison, the proposed

test has the overall p-value 0.0017, which does not depend on the preset significance level. Therefore, it is more powerful in this example.

The second data set is taken from a study on the kidney dialysis patients to assess the time to first exit site infection (in months) in 119 patients with renal impairment. Among those patients, 43 utilized a surgically placed catheter (group 1), 76 utilized a percutaneous placement of their catheter (group 2). There were 27 and 65 censored observations in groups 1 and 2, respectively. This data set was analyzed in the papers by Lin and Wang (Lin and Wang, 2004), and Qiu and Sheng (Qiu and Sheng, 2008). Using the significance level of 0.05, we got the p-values of 0.11 and 0.00078 in the first stage and the second stage of the QS test, and the overall p-value is 0.026. If we change the significance level to 0.001, the overall p-value of the QS test becomes 0.0013, which is larger than 0.001; therefore, we will not reject the null hypothesis. In contrast, the p-value from the proposed test is 0.00090, which is much smaller than the overall p-values of the QS test in both cases.

The third data set was from the randomized, double-blinded Digoxin Intervention Trial (The Digitalis Investigation Group, 1997). In the trial, patients with left ventricular ejection fractions of 0.45 or less were randomly assigned to digoxin (3397 patients) or placebo (3403 patients) groups. A primary outcome was the mortality due to worsening heart failure. The subjects were categorized based on their treatments and gender. It is of interest to compare the survival distributions among the four groups. We applied various methods to this data set. Table 5 lists the p-values obtained by the QS test and the proposed test. Here we have six comparisons in total, using 0.05 as the significance level, a comparison with an adjusted p-value less than 0.05/6, i.e., 0.0083, would be claimed significant. None of the p-values from QS test is less than

0.0083, however, the p-value for comparing the first and second groups from the proposed test is 0.0028. In addition, the p-values from the LR test and the Wilcoxon test are 0.11 and 0.092, respectively.

4. Discussion and Conclusion

In this paper, we proposed an improvement of the QS test by using the Fisher test in defining the overall p-value of the test. There are several advantages to use the proposed method. First, it is more powerful than the original QS test in various cases. Second, unlike the QS test, the p-value of our proposed method is independent of the preset significance level and of the order of the two stages as well. Third, for comparing multiple hazard curves, the proposed test performs much better than the QS test in many cases.

It should be pointed out that in the original QS test, the authors set the default values as $\alpha_1 = \alpha_2 = 1 - \sqrt{1-\alpha}$. We should choose appropriate values for α_1 and α_2 in the TS test if prior information about the hazard curves is available, so that the TS test has optimal detecting power. However, if the values of α_1 and α_2 are set inappropriately, the TS method may also lose power dramatically.

Besides the Fisher test considered here, several other p-value combining methods have been proposed in the literature. For example, the weighted Z tests and the generalized Fisher tests are commonly used in the literature (Chen, 2011; Chen and Nadarajah, 2014; Chen et al., 2014). The method proposed by Chen and Nadarajah (Chen and Nadarajah, 2014) has similar performance to that of the Fisher test. However, the methods based on the weighted Z-test are not recommended since they are not robust to the p-values obtained in the individual steps of the two-stage scheme and can potentially lose power dramatically. In addition, if we have some

prior information about the hazard curves, powerful methods of combining p-values can be designed accordingly.

In the literature, there are many multiple comparison methods, such as the Sidak procedure, Tukey's procedure, and Dunnett's method. However, all of them require some assumptions that may not be valid here. As a comparison, the Bonferroni procedure adopted here does not require any assumptions, such as the independence among individual p-values. Furthermore, our simulation study has shown that the proposed method using the Bonferroni procedure works well in terms of controlling the type I error rate.

It should be pointed out that, although the original QS test and the proposed method are designed for comparing hazard rate functions, they can also be used to compare survival curves. In practice, one may want to compare certain quantiles (e.g., the survival median) (Brookmeyer and Crowley, 1982; Chen, 2014a, b; Chen and Zhang, 2016) instead of the whole survival curves. In such cases, both the QS test and the proposed new method cannot be applied directly, and much future research is needed.

References

Brookmeyer, R., Crowley, J., 1982. A k-sample median test for censored data. J. Amer. Statist. Assoc. 77, 433-440.

Chen, Z., 2011. Is the weighted z-test the best method for combining probabilities from independent tests? J. Evol. Biol. 24, 926-930.

Chen, Z., 2014a. Extension of Mood's median test for survival data. Statistics & Probability Letters 95, 77-84.

Chen, Z., 2014b. A Nonparametric Approach to Detecting the Difference of Survival Medians. Commun Stat Simulat, (in press), DOI: 10.1080/03610918.03612014.03964804.

Chen, Z., Nadarajah, S., 2014. On the optimally weighted z-test for combining probabilities from independent studies. Comput. Stat. Data Anal. 70, 387–394.

Chen, Z., Yang, W., Liu, Q., Yang, J.Y., Li, J., Yang, M.Q., 2014. A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. BMC Bioinformatics 15 (Suppl 17), S3.

Chen, Z., Zhang, G., 2016. Comparing survival curves based on medians. BMC Med. Res. Methodol. 16, 1.

Cheng, M.-Y., Qiu, P., Tan, X., Tu, D., 2009. Confidence intervals for the first crossing point of two hazard functions. Lifetime Data Anal. 15, 441-454.

Fisher, R.A., 1932. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53, 457-481.

Lee, E.T., Wang, J., 2003. Statistical methods for survival data analysis. Wiley-Interscience.

Li, H., Han, D., Hou, Y., Chen, H., Chen, Z., 2015. Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods. PLoS One 10.

Lin, X., Wang, H., 2004. A new testing approach for comparing the overall homogeneity of survival curves. Biometrical Journal 46, 489-496.

Liu, K., Qiu, P., Sheng, J., 2007. Comparing two crossing hazard rates by Cox proportional hazards modelling. Stat. Med. 26, 375-391.

Mantel, N., Bohidar, N.R., Ciminera, J.L., 1977. Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. Cancer Res. 37, 3863-3868.

Mantel, N., Stablein, D.M., 1988. The crossing hazard function problem. The Statistician, 59-64.

Moreau, T., Maccario, J., Lellouch, J., Huber, C., 1992. Weighted log rank statistics for comparing two distributions. Biometrika 79, 195-198.

Owen, A.B., 2009. Karl Pearson's meta-analysis revisited. Ann. Statist 37, 3867–3892.

Park, K.Y., Qiu, P., 2014. Model selection and diagnostics for joint modeling of survival and longitudinal data with crossing hazard rate functions,. Stat. Med. 33, 4532–4546.

Qiu, P., Sheng, J., 2008. A two-stage procedure for comparing hazard rate functions. J. Roy. Stat. Soc. Ser. B. (Stat. Method.) 70, 191-208.

The Digitalis Investigation Group, 1997. The effect of digoxin on mortality and morbidity in patients with heart failure. N. Engl. J. Med. 336, 525-533.

Tables:

Table 1. Empirical type I error rates of each method under different settings with sample size 50 for each group (two groups) and the significance level 0.05. The results were obtained from 1000 replicates.

distribution	method	censoring rate				
		0, 0	0.2,0.2	0.4,0.4	0.6,0.6	0.4,0.6
U: $\theta = (6,6)$ $a = (2,2)$	TS	0.046	0.045	0.043	0.049	0.046
	TS1	0.044	0.042	0.043	0.048	0.045
	TS2	0.049	0.049	0.050	0.052	0.050
	LR	0.053	0.055	0.052	0.052	0.049
	NP	0.039	0.036	0.041	0.046	0.043
	New	0.042	0.040	0.043	0.048	0.046
E: $\theta = (12,12)$ $a = (0.1,0.1)$	TS	0.044	0.045	0.047	0.049	0.044
	TS1	0.043	0.042	0.045	0.048	0.042
	TS2	0.049	0.049	0.054	0.052	0.049
	LR	0.051	0.051	0.055	0.050	0.050
	NP	0.039	0.039	0.039	0.045	0.039
	New	0.040	0.043	0.044	0.046	0.043
LN: $\theta = (5,5)$ $a = (1,1)$	TS	0.047	0.046	0.045	0.050	0.049
	TS1	0.045	0.046	0.040	0.050	0.046
	TS2	0.051	0.049	0.052	0.052	0.053
	LR	0.054	0.050	0.053	0.051	0.053
	NP	0.040	0.042	0.035	0.046	0.043
	New	0.045	0.040	0.041	0.051	0.046

Table 2. Empirical powers of each method under different settings with sample size 50 for each group (two groups) and the significance level 0.05. The results were obtained from 1000 replicates.

distribution	method	censoring rate				
		0, 0	0.2,0.2	0.4,0.4	0.6,0.6	0.4,0.6
U: $\theta = (6,6)$ $a = (8,3)/3$	TS	0.616	0.547	0.4750	0.375	0.388
	TS1	0.606	0.607	0.550	0.445	0.465
	TS2	0.549	0.444	0.355	0.298	0.287
	LR	0.243	0.105	0.049	0.067	0.068
	NP	0.569	0.605	0.570	0.469	0.488
	New	0.769	0.586	0.463	0.381	0.367
E: $\theta = (12,10)$ $a = (4,5)/40$	TS	0.730	0.743	0.749	0.774	0.780
	TS1	0.720	0.734	0.740	0.757	0.783
	TS2	0.709	0.721	0.728	0.751	0.757
	LR	0.593	0.595	0.600	0.631	0.639
	NP	0.632	0.652	0.637	0.614	0.729
	New	0.810	0.825	0.832	0.846	0.854
LN: $\theta = (5,5.6)$ $a = (1,1)$	TS	0.723	0.676	0.604	0.543	0.575
	TS1	0.635	0.585	0.514	0.447	0.468
	TS2	0.762	0.721	0.648	0.598	0.630

	LR	0.773	0.733	0.666	0.619	0.642
	NP	0.265	0.243	0.204	0.145	0.120
	New	0.748	0.708	0.625	0.560	0.581
U+E: $\theta = (10, 4)$ $a = (5,0.1)$	TS	0.908	0.607	0.544	0.526	0.659
	TS1	0.832	0.553	0.577	0.593	0.700
	TS2	0.927	0.622	0.456	0.409	0.547
	LR	0.899	0.485	0.161	0.069	0.115
	NP	0.156	0.342	0.557	0.613	0.715
	New	0.991	0.789	0.592	0.520	0.658
U+LN: $\theta = (5, 1)$ $a = (5,1)$	TS	0.733	0.449	0.352	0.404	0.375
	TS1	0.793	0.462	0.296	0.303	0.295
	TS2	0.645	0.419	0.383	0.471	0.419
	LR	0.163	0.287	0.393	0.490	0.440
	NP	0.800	0.422	0.136	0.049	0.086
	New	0.752	0.505	0.384	0.388	0.366
E+LN: $\theta = (0, 1)$ $a = (0.2, 2)$	TS	0.667	0.525	0.401	0.269	0.301
	TS1	0.667	0.582	0.467	0.306	0.351
	TS2	0.602	0.425	0.296	0.213	0.228
	LR	0.247	0.097	0.053	0.096	0.080
	NP	0.616	0.588	0.491	0.308	0.369
	New	0.864	0.538	0.369	0.278	0.297

Table 3. Empirical type I error rates of each method under different settings with sample size 50 for each group (four groups) and the significance level 0.05. The results were obtained from 1000 replicates.

distribution	method	censoring rate				
		0, 0,0,0	(2,2,2,2)/10	(4,4,4,4)/10	(6,6,6,6)/10	(4,6,4,6)/10
U: $\theta = (6,6,6,6)$ $a = (2,2,2,2)$	TS	0.046	0.040	0.040	0.044	0.048
	TS1	0.046	0.040	0.040	0.044	0.048
	TS2	0.046	0.040	0.040	0.044	0.048
	LR	0.059	0.057	0.049	0.055	0.049
	New	0.030	0.031	0.035	0.044	0.043
E: $\theta = (12,12,12,12)$ $a = (0.1,0.1,0.1,0.1)$	TS	0.044	0.035	0.048	0.043	0.044
	TS1	0.044	0.035	0.048	0.043	0.044
	TS2	0.044	0.035	0.048	0.043	0.044
	LR	0.053	0.046	0.059	0.063	0.055
	New	0.029	0.028	0.045	0.050	0.035
LN: $\theta = (5,5,5,5)$ $a = (1,1,1,1)$	TS	0.060	0.052	0.047	0.058	0.051
	TS1	0.060	0.052	0.047	0.058	0.051
	TS2	0.060	0.052	0.047	0.058	0.051
	LR	0.060	0.051	0.076	0.073	0.054
	New	0.032	0.037	0.040	0.050	0.052

Table 4. Empirical powers of each method under different settings with sample size 50 for each group (four groups) and the significance level 0.05. The results were obtained from 1000 replicates.

distribution	method	censoring rate				
		0, 0,0,0	(2,2,2,2)/10	(4,4,4,4)/10	(6,6,6,6)/10	(4,6,4,6)/10

U: $\theta = (6,6,6,6)$ $a = (8,8,3,3)/3$	TS	0.676	0.380	0.069	0.286	0.217
	TS1	0.676	0.380	0.069	0.286	0.217
	TS2	0.676	0.380	0.069	0.286	0.217
	LR	0.823	0.507	0.092	0.329	0.149
	New	1.000	1.000	1.000	0.994	0.999
E: $\theta = (12,10,12,10)$ $a = (4,4,5,5)/40$	TS	0.412	0.418	0.424	0.484	0.502
	TS1	0.412	0.418	0.424	0.484	0.502
	TS2	0.412	0.418	0.424	0.484	0.502
	LR	0.441	0.449	0.476	0.554	0.557
	New	0.736	0.789	0.802	0.842	0.828
LN: $\theta = (5,5,5.6,5.6)$ $a = (1,1,1,1)$	TS	0.850	0.814	0.773	0.687	0.725
	TS1	0.850	0.814	0.773	0.687	0.725
	TS2	0.850	0.814	0.773	0.687	0.725
	LR	0.913	0.876	0.857	0.765	0.803
	New	0.854	0.814	0.767	0.654	0.716
U+E: $\theta = (10,10,4,4)$ $a = (5,5,0.1,0.1)$	TS	0.931	0.483	0.135	0.063	0.108
	TS1	0.931	0.483	0.135	0.063	0.108
	TS2	0.931	0.483	0.135	0.063	0.108
	LR	0.976	0.602	0.192	0.076	0.142
	New	0.933	0.537	0.509	0.526	0.537
U+LN: $\theta = (5,5,1,1)$ $a = (5,5,1,1)$	TS	0.166	0.293	0.445	0.550	0.511
	TS1	0.166	0.293	0.445	0.550	0.511
	TS2	0.166	0.293	0.445	0.550	0.511
	LR	0.194	0.342	0.497	0.628	0.537
	New	0.842	0.602	0.437	0.470	0.455
E+LN: $\theta = (0,0,1,1)$ $a = (0.2,0.2,2,2)$	TS	0.198	0.079	0.039	0.113	0.065
	TS1	0.198	0.079	0.039	0.113	0.065
	TS2	0.198	0.079	0.039	0.113	0.065
	LR	0.310	0.108	0.049	0.126	0.075
	New	0.610	0.471	0.369	0.291	0.352

Table 5. P-values obtained by the QS test and the proposed method from each pair of groups of the DIT data.

Group pair	LR	NP	TS	New
1 vs. 2	0.019	0.026	0.019	0.0028
1 vs. 3	0.17	0.32	0.36	0.21
1 vs. 4	0.38	0.16	0.18	0.23
2 vs. 3	0.86	0.016	0.041	0.073
2 vs. 4	0.47	0.012	0.037	0.035
3 vs. 4	0.66	0.69	0.70	0.81

Figure Legend:

Figure 1. The acceptance regions for the original QS two-stage test and the proposed method based on the Fisher test.