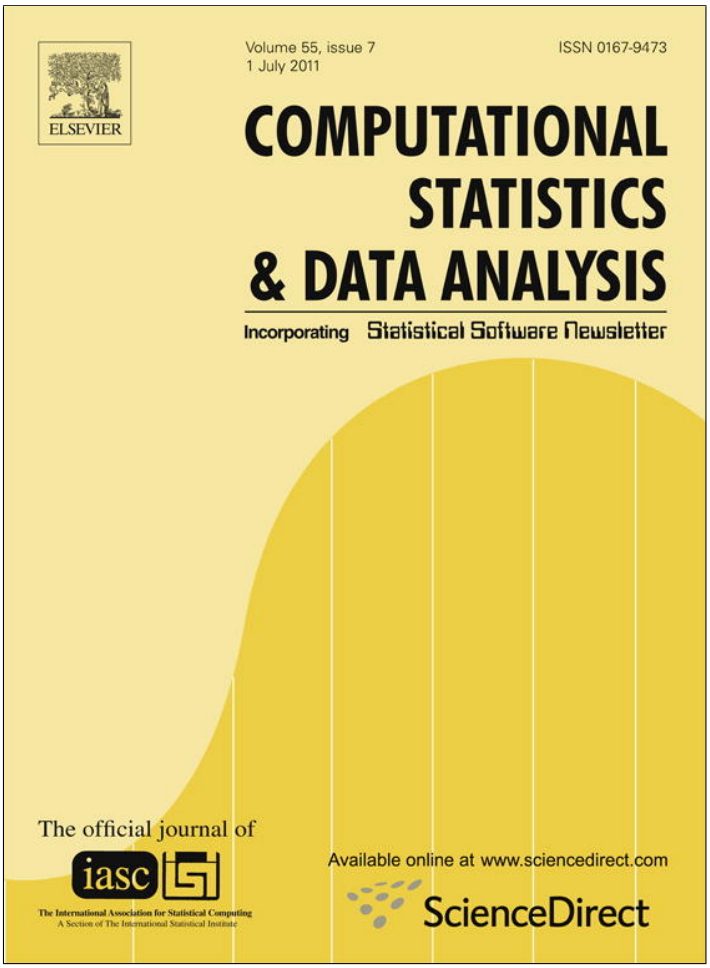

Computational Statistics and Data Analysis 55 (2011) 2400–2409

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A pretest for using logrank or Wilcoxon in the two-sample problem

Annie Tordilla Darilay *, Joshua D. Naranjo

Department of Statistics, Western Michigan University, Kalamazoo, MI 49009, USA

ARTICLE INFO

Article history:
Received 5 May 2010
Received in revised form 3 February 2011
Accepted 4 February 2011
Available online 12 February 2011

Keywords:
Survival analysis
Logrank
Wilcoxon
Adaptive test
Accelerated failure time

ABSTRACT

In a two-sample location-scale model with censored data, the logrank test is asymptotically efficient when the error distribution is extreme minimum value. On the other hand, the Wilcoxon test is asymptotically efficient when the error distribution is logistic. We propose a pretest for choosing between logrank and Wilcoxon by determining if the error distribution is closer to extreme minimum value or logistic. This adaptive test is compared with the logrank and Wilcoxon tests through simulation.

1. Introduction

In survival analysis, treatment efficacy is often analyzed by comparing the survival rates of two treatment groups. Two commonly used tests for the comparison of survival distributions are the generalized Wilcoxon procedure (Gehan, 1965; Peto and Peto, 1972) and the logrank test (Mantel, 1966; Cox, 1972; Peto and Peto, 1972). Both tests are based on the ranks of the observations, and have several versions in the literature. In this paper, the Wilcoxon and logrank tests will refer to Peto and Peto's (1972) version of the statistics proposed by Gehan (1965) and Mantel (1966), respectively. Leton and Zuluaga (2005) present a comprehensive summary of the different names, versions and representations of the generalized Wilcoxon and logrank tests.

The finite sample performance of these tests have been compared in several simulation studies. Lee et al. (1975) compared size and power of the tests using small samples from exponential and Weibull survival distributions with and without censoring. Latta (1981) extended the simulations to include the lognormal survival distribution, allow for unequal sample sizes, and allow for censoring in only one sample. Beltangady and Frankowski (1989) focussed on the effect of unequal censoring, using various combinations of censoring proportions. More recently, Leton and Zuluaga (2001, 2005) compared the performance of various versions of the generalized Wilcoxon and logrank tests under scenarios of early and late hazard differences.

In general, the logrank test is most powerful when the ratio of the two hazard functions is constant. When hazard ratio is nonconstant, the generalized Wilcoxon test may perform better than the logrank test especially when differences occur early in time (Lee et al., 1975; Leton and Zuluaga, 2001, 2005). In applications, the logrank test is often used after checking for validity of the proportional hazards (PH) assumption, with Wilcoxon being the fallback method when the PH assumption fails. There have been several graphical methods suggested for assessing the proportional hazards assumption (Hess, 1995). One commonly used graphical method that is available on many statistical software (i.e. SAS, Stata and R) is the plotting of the log of the cumulative hazard function against log time and checking for parallelism.

* Corresponding address: 12 Coolpond Ct, Baltimore, MD 21227, USA. Tel.: +1 269 267 1513.
E-mail addresses: annie_darilay@yahoo.com (A.T. Darilay), joshua.naranjo@wmich.edu (J.D. Naranjo).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.02.006

Table 1
Distributions of Y and e^Y.

m	Distribution of Y	Distribution of e^Y
1	Logistic	Loglogistic
∞	Extreme minimum value	Weibull

The properties of the logrank and Wilcoxon tests are discussed extensively in the literature (Breslow, 1970; Cox, 1972; Peto, 1972; Peto and Peto, 1972; Kalbfleisch and Prentice, 1980; Andersen et al., 1993; Klein and Moeschberger, 2003). Under a location-scale model, $y = \mu + \beta z + \sigma\epsilon$ where y is right censored, the logrank test is known to be fully efficient for the extreme minimum value error distribution $f(\epsilon) = \exp(\epsilon - e^{\epsilon})$, while Peto and Peto's Wilcoxon test is fully efficient for the logistic error distribution $f(\epsilon) = e^{\epsilon}/(1 + e^{\epsilon})^2$ (Prentice, 1978; Gill, 1980). Since the efficiency of rank tests is invariant under monotone increasing data transformations, this also means that the logrank test is fully efficient when the failure time follows a Weibull distribution and Peto and Peto's Wilcoxon test is fully efficient when the failure time follows a loglogistic distribution.

Prentice (1975) discussed a general family of distributions that contain both the extreme minimum value and logistic distributions. Maximum likelihood estimates of parameters can be used as a diagnostic for discriminating between the extreme minimum value and logistic in the context of one sample. In this paper, we extend his procedure to the two-sample problem and propose a statistic for detecting whether the common error distribution is closer to extreme minimum value or logistic. It may be used as a pretest for logrank validity, or as the first part of an adaptive test which uses logrank or Wilcoxon depending on whether the underlying distribution seems closer to extreme minimum value or logistic. Simulation results show that an adaptive test will have efficiency closer to the better of the two tests under either extreme minimum value or logistic. The Type I error rate is inflated minimally.

Section 2 contains details on distributions and maximum likelihood estimation. Section 3 describes a proposal for calculating the statistic in the two-sample problem, and using it to adapt between logrank and Wilcoxon. Simulation results for the adaptive test are presented in Section 4 and application of the test to a real data set is provided in Section 5.

2. Weibull versus loglogistic distribution

Discriminating between Weibull and loglogistic distributions under the accelerated failure time model is equivalent to discriminating between their error distributions: extreme minimum value and logistic distributions. In this section, we will present Prentice's (1975) discrimination procedure that embeds both extreme minimum value and logistic distributions in a larger parametric family of distributions.

2.1. Parametric family of distributions

Consider the location-scale model, $y = \mu + \sigma\epsilon$, such that

$$f(\epsilon) = \frac{1}{mB(1,m)} e^{\epsilon}\left(1 + \frac{e^{\epsilon}}{m}\right)^{-(m+1)} \tag{1}$$

where $m > 0$ and B is the beta function $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\partial t$. The density of Y is

$$f(y) = \frac{1}{\sigma mB(1,m)} e^{\left(\frac{y-\mu}{\sigma}\right)}\left(1 + \frac{e^{\left(\frac{y-\mu}{\sigma}\right)}}{m}\right)^{-(m+1)}. \tag{2}$$

The distribution in (2) is a subset of a larger family of distributions discussed in Prentice (1975). Table 1 lists special cases of the distribution of Y and $\exp(Y)$.

Since the logistic and extreme minimum value distributions correspond to $m = 1$ and $m = \infty$, respectively, a diagnostic for choosing between these two models can be based on an estimate of m being closer to 1 or ∞.

2.2. Maximum likelihood estimation of m

Let $y_1, y_2, \ldots, y_n$ be a random sample from the distribution (2). The likelihood of the sample is

$$L(y;\mu,\sigma,m) = \left(\frac{1}{\sigma mB(1,m)}\right)^n e^{\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)} \prod_{i=1}^{n}\left[1 + \frac{e^{\left(\frac{y_i-\mu}{\sigma}\right)}}{m}\right]^{-(m+1)}. \tag{3}$$

Table 2
Percentiles of $\hat{m}$ at $\mu = 0$.

Percentile	$\sigma = 0.5$		$\sigma = 2$	
	Extreme minimum value ($m = \infty$)	Logistic ($m = 1$)	Extreme minimum value ($m = \infty$)	Logistic ($m = 1$)
5th	2.2716	0.4310	2.1767	0.4418
10th	3.1288	0.5283	3.2156	0.5402
20th	5.8244	0.6590	6.0527	0.6724
30th	13.7306	0.7733	13.5070	0.7921
40th	3 768.6133	0.8871	3 523.0438	0.9048
50th	18 159.3947	1.0165	40 544.4337	1.0444
60th	25 137.5983	1.1816	61 759.8040	1.2092
70th	30 609.9432	1.4183	75 824.0997	1.4498
80th	36 062.4104	1.7928	89 861.8513	1.8903
90th	42 404.6878	3.0453	106 864.0991	3.1853
95th	47 324.8637	6.1804	120 990.5581	6.1366

Taking partial derivatives of the log-likelihood and equating to 0, we get the following estimating equations

$$\frac{\partial \ln L}{\partial \mu} = -\frac{n}{\sigma} + \frac{(1+m)}{\sigma}\sum_{i=1}^{n}\frac{1}{1+\frac{1}{m}e^{\epsilon_i}}\left(\frac{e^{\epsilon_i}}{m}\right) = 0, \tag{4}$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{\sum_{i=1}^{n}\epsilon_i}{\sigma} + \frac{(1+m)}{\sigma}\sum_{i=1}^{n}\frac{\epsilon_i}{1+\frac{1}{m}e^{\epsilon_i}}\left(\frac{e^{\epsilon_i}}{m}\right) - \frac{n}{\sigma} = 0, \tag{5}$$

$$\frac{\partial \ln L}{\partial m} = -\frac{n}{m} + (1+m)\sum_{i=1}^{n}\frac{e^{\epsilon_i}}{m^2+me^{\epsilon_i}} - \sum_{i=1}^{n}\ln\left(1+\frac{1}{m}e^{\epsilon_i}\right) - n[\psi(m) - \psi(1+m)] = 0 \tag{6}$$

where $\epsilon_i = (y_i - \mu)/\sigma$. To obtain (6), note that $\partial B(a,b)/\partial b = B(a,b)[\psi(b) - \psi(a+b)]$, where $\psi(k) = \partial \ln \Gamma(k)/\partial(k) = \int_0^\infty [e^{-t}t^{-1} - e^{-kt}(1-e^{-t})^{-1}]\partial t$ is the digamma function (Davis, 1965).

The MLE $\hat{m}$ can be obtained simultaneously with $\hat{\mu}$ and $\hat{\sigma}$ using numerical methods. For our simulation, we will use R's *optim* function which outputs the three maximum likelihood estimates based on the inputted log-likelihood function and the estimating Eqs. (4)–(6).

To confirm that $\hat{m}$ can indeed discriminate between the extreme minimum value and logistic distributions, we generated 50 observations from either distribution and estimated $\hat{m}$. The simulation was replicated 5000 times for each of the following cases:

CaseA. $\mu = 0, \sigma < 1$ (pdf of both distributions is narrow and tall)
CaseB. $\mu = 0, \sigma > 1$ (pdf of both distributions is broad and shallow).

Observe from Table 2 that $\hat{m}$ tends to be small under the logistic distribution (80% are smaller than 2.0), and tends to be quite large under the extreme minimum value distribution (95% are larger than 2.0).

The asymptotic distribution of $\hat{m}$ can be derived using standard maximum likelihood theory. More precisely, we present the asymptotic distribution of $1/\hat{m}$, which has finite mean. If $m = 1$,

$$\frac{1}{\hat{m}} \sim N\left\{1, \frac{4}{n}\left(\frac{\pi^2+3}{\pi^2-6}\right)\right\}, \tag{7}$$

approximately when n is large. If $m = \infty$,

$$\frac{1}{\hat{m}} \sim N\left\{0, \frac{1}{n\left(1-\frac{6}{\pi^2}\right)}\right\} \tag{8}$$

approximately. See Prentice (1975) for details. The asymptotic variances above evaluate to $13.30/n$ and $2.55/n$, respectively.

Eqs. (7) and (8) and Table 2 suggest a pretest for logistic or extreme minimum value depending on whether $1/\hat{m} > 0.5$, or equivalently, $\hat{m} < 2.0$.

In Sections 3 and 4, we investigate the properties of an adaptive test which uses Wilcoxon when $\hat{m} < 2.0$ and logrank otherwise.

3. Adaptive survival test

Let $T_1 = (T_{11}, T_{12}, \ldots, T_{1n_1})$ and $T_2 = (T_{21}, T_{22}, \ldots, T_{2n_2})$ be random samples of failure times whose log values are denoted by the following respective location-scale models:

$$\log T_1 = \mu + \sigma\epsilon \tag{9}$$

$$\log T_2 = \mu + \beta + \sigma\epsilon, \tag{10}$$

where the distribution of ϵ is unknown.

Based on (9) and (10), the relationship between $\log T_1$ and $\log T_2$ is characterized by the location-change model

$$\log T_2 = \beta + \log T_1$$

while the relationship between T_1 and T_2 is characterized by the scale-change model

$$T_2 = e^{\beta} T_1. \tag{11}$$

In testing $\beta = 0$, it is known that the logrank test is asymptotically efficient for the extreme minimum value error distribution while Peto–Peto's Wilcoxon test is asymptotically efficient for the logistic error distribution (Prentice, 1978). Now we can use the diagnostic $\hat{m}$ on the combined log failure times to decide whether the extreme minimum value or logistic distribution better fits the data. However, the two samples cannot be combined before they are standardized to the same scale.

From model (11), the scale-change parameter or acceleration factor is $e^{\beta} = T_2/T_1$, which can be intuitively estimated by the ratio of the two standard deviations, i.e. $SD(T_2)/SD(T_1)$. Hence, multiplying T_1 by this ratio,

$$T_1^{\star} = \frac{SD(T_2)}{SD(T_1)} T_1,$$

results in a rescaled $T_1^{\star}$ which exhibits the same variability as T_2.

We now propose an adaptive test procedure for $H_0 : \beta = 0$. Let T_1 and C_1 be the failure and censored times of the first sample D_1. Let T_2 and C_2 be the failure and censored times of the second sample D_2.

a. Standardize the failure times from the two samples so they have the same variability:

$$T_1^{*} = T_1 \times \frac{SD(T_2)}{SD(T_1)}.$$

b. Combine the standardized failure times from the two samples $T_c = (T_1^{*}$ and $T_2)$.
c. Log-transform T_c and compute $\hat{m}$.
d. If $\hat{m} < 2.0$, use the Wilcoxon test on D_1 and D_2. Otherwise, use the logrank test.

We chose the Peto–Peto version of the logrank and Wilcoxon tests. Unlike the Gehan version, Peto–Peto's Wilcoxon test does not depend on the censoring rates and thus does not give misleading results when the censoring patterns are different (Prentice and Marek, 1979).

4. Simulation

In this section, we will examine the finite sample performance of the adaptive survival test for the two-sample accelerated failure time problem. We will assess its empirical validity, and compare its power against the dedicated Wilcoxon and logrank tests, respectively.

Failure times, T_1 and T_2, were generated from the following distributions:

a. Weibull

$$f(t) = \frac{1}{\sigma} e^{(-\mu/\sigma)} t^{(1/\sigma - 1)} \exp[-e^{(-\mu/\sigma)} t^{(1/\sigma)}], \quad \sigma > 0.$$

b. Loglogistic

$$f(t) = \frac{\frac{1}{\sigma} e^{(-\mu/\sigma)} t^{(1/\sigma - 1)}}{\left[1 + e^{(-\mu/\sigma)} t^{(1/\sigma)}\right]^2}, \quad \sigma > 0.$$

c. Lognormal

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-0.5[\ln(t) - \mu]^2 \sigma^{-2}}, \quad \sigma > 0.$$

For each of these failure time distributions, the shape parameter σ is set to (i) greater than 1, (ii) equal to 1, and (iii) less than 1 to represent different degrees of skewness and tail weight of the survival function. The scale parameter for S_1 is $\mu = 0$ in all cases. In the size study, the scale parameter for S_2 is also $\mu = 0$ while in the power study, the scale parameter for S_2 is $\mu > 0$.

The simulations were also done using different levels of right censoring. The censoring distributions of the two groups were equal and uniform $U(0, c)$. The values of c were used to control the desired censoring percentages considered in the simulation.

Table 3
Size simulation results at significance level $\alpha = 0.05$, from 10,000 replications with equal censoring, $n_1 = n_2 = 20$.

Relative skewness	Survival time	Average % censored	Adaptive	Logrank	Wilcoxon
$S(t)$ = Weibull distribution					
		$C_1 = 0.0, C_2 = 0.0$	0.0599	0.0586	0.0522
Heavy	$\mu_1 = 0, \sigma_1 = 2$	$C_1 = 23.2, C_2 = 23.4$	0.0519	0.0519	0.0505
	$\mu_2 = 0, \sigma_2 = 2$	$C_1 = 29.7, C_2 = 29.7$	0.0540	0.0538	0.0503
		$C_1 = 0.0, C_2 = 0.0$	0.0596	0.0583	0.0560
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 16.6, C_2 = 16.6$	0.0558	0.0556	0.0520
	$\mu_2 = 0, \sigma_2 = 1$	$C_1 = 24.3, C_2 = 24.6$	0.0558	0.0560	0.0511
		$C_1 = 0.0, C_2 = 0.0$	0.0613	0.0600	0.0559
Light	$\mu_1 = 0, \sigma_1 = 0.75$	$C_1 = 15.3, C_2 = 15.3$	0.0571	0.0564	0.0514
	$\mu_2 = 0, \sigma_2 = 0.75$	$C_1 = 22.8, C_2 = 22.8$	0.0565	0.0558	0.0549
$S(t)$ = Loglogistic distribution					
		$C_1 = 0.0, C_2 = 0.0$	0.0678	0.0594	0.0538
Heavy	$\mu_1 = 0, \sigma_1 = 1.5$	$C_1 = 22.4, C_2 = 22.3$	0.0571	0.0572	0.0523
	$\mu_2 = 0, \sigma_2 = 1.5$	$C_1 = 35.3, C_2 = 35.4$	0.0523	0.0525	0.0503
		$C_1 = 0.0, C_2 = 0.0$	0.0678	0.0587	0.0507
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 15.1, C_2 = 15.2$	0.0526	0.0515	0.0472
	$\mu_2 = 0, \sigma_2 = 1$	$C_1 = 29.6, C_2 = 29.5$	0.0544	0.0533	0.0510
		$C_1 = 0.0, C_2 = 0.0$	0.0658	0.0561	0.0517
Light	$\mu_1 = 0, \sigma_1 = 0.5$	$C_1 = 7.6, C_2 = 7.6$	0.0633	0.0570	0.0517
	$\mu_2 = 0, \sigma_2 = 0.5$	$C_1 = 20.4, C_2 = 20.5$	0.0578	0.0545	0.0484
$S(t)$ = Lognormal distribution					
		$C_1 = 0.0, C_2 = 0.0$	0.0653	0.0574	0.0531
Heavy	$\mu_1 = 0, \sigma_1 = 1.5$	$C_1 = 16.3, C_2 = 16.4$	0.0587	0.0567	0.0508
	$\mu_2 = 0, \sigma_2 = 1.5$	$C_1 = 31.2, C_2 = 31.1$	0.0557	0.0556	0.0522
		$C_1 = 0.0, C_2 = 0.0$	0.0690	0.0608	0.0527
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 10.9, C_2 = 10.8$	0.0605	0.0555	0.0505
	$\mu_2 = 0, \sigma_2 = 1$	$C_1 = 25.2, C_2 = 25.4$	0.0612	0.0595	0.0565
		$C_1 = 0.0, C_2 = 0.0$	0.0639	0.0564	0.0513
Light	$\mu_1 = 0, \sigma_1 = 0.75$	$C_1 = 8.9, C_2 = 8.9$	0.0645	0.0591	0.0509
	$\mu_2 = 0, \sigma_2 = 0.75$	$C_1 = 21.9, C_2 = 21.8$	0.0603	0.0561	0.0499

4.1. Size study

Each test was conducted on 10,000 samples under each survival and censoring configuration. Results at the 5% significance level for sample sizes 20 and 50 are presented in Tables 3 and 4, respectively. All simulations were done in R.

At $n = 20$, all 3 tests seem to be a touch liberal, with the size of the adaptive test slightly larger than either non-adaptive test (though not by much). The Wilcoxon test tends to be the most conservative in almost all cases.

At $n = 50$, the observed significance level of all tests are closer to the nominal value of 0.05, with the Wilcoxon again being the most conservative.

4.2. Power study

The power of the tests at a 5% significance level was obtained from 10,000 samples for each survival and censoring configuration. The results for Weibull, loglogistic, and lognormal are displayed on Tables 5–7.

The logrank test is known to be efficient under the Weibull survival distribution. Our simulations confirm this, with logrank outperforming Wilcoxon in all cases. The difference in their power is minimal under heavily skewed Weibull distributions with censoring but increases up to 10% as the degree of skewness decreases. The adaptive test approximates the power of the logrank test, and is better than Wilcoxon in this situation.

The simulation results support the known result that Wilcoxon is efficient under the loglogistic survival distribution. Wilcoxon beats logrank by up to 10% in the uncensored cases and up to 4% in the censored cases. At $n = 50$, all tests perform well under lightly skewed survival distributions, with Wilcoxon beating logrank by 1% to 2% efficiency. The adaptive test approximates the power of the efficient Wilcoxon in the uncensored cases. Under censoring, the adaptive test is either better than or comparable to logrank.

Under the lognormal survival distribution, Wilcoxon beats logrank by 2% to 6% efficiency, depending on the degree of skewness of the survival distribution and the censoring rate. The adaptive test is more powerful than logrank. Generally, as the censoring rate decreases or as the survival distribution becomes less skewed, the gain in power of the adaptive test against logrank increases.

As expected, the power of all three tests decrease with increasing skewness of the survival distribution and increasing percentage of censored observations. Between the logrank and Wilcoxon tests, the logrank test is better under the Weibull survival distribution while the Wilcoxon does better under the loglogistic distribution. When the distribution is lognormal,

Table 4
Size simulation results at significance level $\alpha = 0.05$, from 10,000 replications with equal censoring, $n_1 = n_2 = 50$.

Relative skewness	Survival time	Average % censored	Adaptive	Logrank	Wilcoxon
$S(t)$ = Weibull distribution					
		$C_1 = 0.0, C_2 = 0.0$	0.0545	0.0542	0.0485
Heavy	$\mu_1 = 0, \sigma_1 = 2$	$C_1 = 29.8, C_2 = 29.7$	0.0552	0.0552	0.0556
	$\mu_2 = 0, \sigma_2 = 2$	$C_1 = 41.4, C_2 = 41.3$	0.0489	0.0489	0.0484
		$C_1 = 0.0, C_2 = 0.0$	0.0556	0.0522	0.0496
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 16.6, C_2 = 16.4$	0.0540	0.0541	0.0510
	$\mu_2 = 0, \sigma_2 = 1$	$C_1 = 43.3, C_2 = 43.3$	0.0502	0.0502	0.0488
		$C_1 = 0.0, C_2 = 0.0$	0.0556	0.0556	0.0487
Light	$\mu_1 = 0, \sigma_1 = 0.75$	$C_1 = 22.9, C_2 = 23.0$	0.0477	0.0479	0.0480
	$\mu_2 = 0, \sigma_2 = 0.75$	$C_1 = 43.9, C_2 = 43.7$	0.0501	0.0501	0.0503
$S(t)$ = Loglogistic distribution					
		$C_1 = 0.0, C_2 = 0.0$	0.0569	0.0529	0.0498
Heavy	$\mu_1 = 0, \sigma_1 = 1.5$	$C_1 = 30.5, C_2 = 30.5$	0.0513	0.0512	0.0476
	$\mu_2 = 0, \sigma_2 = 1.5$	$C_1 = 43.4, C_2 = 43.5$	0.0519	0.0520	0.0491
		$C_1 = 0.0, C_2 = 0.0$	0.0597	0.0530	0.0485
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 24.0, C_2 = 24.0$	0.0525	0.0526	0.0518
	$\mu_2 = 0, \sigma_2 = 1$	$C_1 = 40.4, C_2 = 40.4$	0.0483	0.0485	0.0462
		$C_1 = 0.0, C_2 = 0.0$	0.0581	0.0549	0.0474
Light	$\mu_1 = 0, \sigma_1 = 0.5$	$C_1 = 14.7, C_2 = 14.8$	0.0555	0.0543	0.0514
	$\mu_2 = 0, \sigma_2 = 0.5$	$C_1 = 33.2, C_2 = 33.1$	0.0526	0.0531	0.0510
$S(t)$ = Lognormal distribution					
		$C_1 = 0.0, C_2 = 0.0$	0.0592	0.0530	0.0505
Heavy	$\mu_1 = 0, \sigma_1 = 1.5$	$C_1 = 25.8, C_2 = 25.8$	0.0519	0.0516	0.0525
	$\mu_2 = 0, \sigma_2 = 1.5$	$C_1 = 39.6, C_2 = 39.4$	0.0563	0.0559	0.0552
		$C_1 = 0.0, C_2 = 0.0$	0.0606	0.0559	0.0498
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 19.5, C_2 = 19.6$	0.0541	0.0529	0.0491
	$\mu_2 = 0, \sigma_2 = 1$	$C_1 = 35.1, C_2 = 35.0$	0.0583	0.0578	0.0546
		$C_1 = 0.0, C_2 = 0.0$	0.0564	0.0519	0.0469
Light	$\mu_1 = 0, \sigma_1 = 0.75$	$C_1 = 16.5, C_2 = 16.5$	0.0579	0.0524	0.0503
	$\mu_2 = 0, \sigma_2 = 0.75$	$C_1 = 32.0, C_2 = 32.0$	0.0555	0.0535	0.0533

Table 5
Power simulation results for Weibull distribution at significance level $\alpha = 0.05$, from 10,000 replications with equal censoring.

Relative skewness	Survival time	Average % censored	Adaptive	Logrank	Wilcoxon
$n_1 = n_2 = 20$					
		$C_1 = 0.0, C_2 = 0.0$	0.2158	0.2141	0.1736
Heavy	$\mu_1 = 0, \sigma_1 = 2$	$C_1 = 23.5, C_2 = 35.4$	0.1655	0.1655	0.1485
	$\mu_2 = 0.75, \sigma_2 = 2$	$C_1 = 29.6, C_2 = 42.2$	0.1575	0.1574	0.1482
		$C_1 = 0.0, C_2 = 0.0$	0.6258	0.6252	0.5222
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 16.8, C_2 = 33.2$	0.5122	0.5137	0.4469
	$\mu_2 = 0.75, \sigma_2 = 1$	$C_1 = 24.5, C_2 = 44.9$	0.4630	0.4644	0.4083
		$C_1 = 0.0, C_2 = 0.0$	0.8500	0.8485	0.7489
Light	$\mu_1 = 0, \sigma_1 = 0.75$	$C_1 = 15.4, C_2 = 32.1$	0.7535	0.7546	0.6688
	$\mu_2 = 0.75, \sigma_2 = 0.75$	$C_1 = 22.9, C_2 = 45.9$	0.6990	0.7003	0.6294
$n_1 = n_2 = 50$					
		$C_1 = 0.0, C_2 = 0.0$	0.2329	0.2324	0.1897
Heavy	$\mu_1 = 0, \sigma_1 = 2$	$C_1 = 29.7, C_2 = 38.1$	0.1713	0.1713	0.1611
	$\mu_2 = 0.5, \sigma_2 = 2$	$C_1 = 41.4, C_2 = 49.7$	0.1527	0.1527	0.1440
		$C_1 = 0.0, C_2 = 0.0$	0.6853	0.6851	0.5648
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 24.6, C_2 = 37.7$	0.5344	0.5344	0.4698
	$\mu_2 = 0.5, \sigma_2 = 1$	$C_1 = 43.4, C_2 = 57.9$	0.4027	0.4027	0.3840
		$C_1 = 0.0, C_2 = 0.0$	0.8970	0.8969	0.8061
Light	$\mu_1 = 0, \sigma_1 = 0.75$	$C_1 = 22.9, C_2 = 37.1$	0.7756	0.7758	0.7028
	$\mu_2 = 0.5, \sigma_2 = 0.75$	$C_1 = 43.7, C_2 = 61.6$	0.6203	0.6208	0.5802

Table 6
Power simulation results for loglogistic distribution at significance level $\alpha = 0.05$, from 10,000 replications with equal censoring.

Relative skewness	Survival time	Average % censored	Adaptive	Logrank	Wilcoxon
$n_1 = n_2 = 20$					
		$C_1 = 0.0, C_2 = 0.0$	0.2340	0.1980	0.2233
Heavy	$\mu_1 = 0, \sigma_1 = 1.5$	$C_1 = 22.5, C_2 = 34.6$	0.2017	0.2001	0.2119
	$\mu_2 = 1, \sigma_2 = 1.5$	$C_1 = 35.2, C_2 = 50.1$	0.2044	0.2043	0.2068
		$C_1 = 0.0, C_2 = 0.0$	0.4337	0.3802	0.4270
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 15.3, C_2 = 28.9$	0.3872	0.3825	0.4116
	$\mu_2 = 1, \sigma_2 = 1$	$C_1 = 29.6, C_2 = 49.3$	0.3792	0.3776	0.3924
		$C_1 = 0.0, C_2 = 0.0$	0.9165	0.8788	0.9378
Light	$\mu_1 = 0, \sigma_1 = 0.5$	$C_1 = 7.6, C_2 = 19.6$	0.8940	0.8758	0.9236
	$\mu_2 = 1, \sigma_2 = 0.5$	$C_1 = 20.2, C_2 = 46.6$	0.8712	0.8655	0.8979
$n_1 = n_2 = 50$					
		$C_1 = 0.0, C_2 = 0.0$	0.3323	0.2790	0.3301
Heavy	$\mu_1 = 0, \sigma_1 = 1.5$	$C_1 = 30.6, C_2 = 41.7$	0.2967	0.2965	0.3142
	$\mu_2 = 0.8, \sigma_2 = 1.5$	$C_1 = 43.5, C_2 = 55.5$	0.2716	0.2715	0.2841
		$C_1 = 0.0, C_2 = 0.0$	0.6114	0.5200	0.6210
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 23.9, C_2 = 38.0$	0.5502	0.5502	0.5896
	$\mu_2 = 0.8, \sigma_2 = 1$	$C_1 = 40.2, C_2 = 57.3$	0.5251	0.5251	0.5459
		$C_1 = 0.0, C_2 = 0.0$	0.9893	0.9758	0.9943
Light	$\mu_1 = 0, \sigma_1 = 0.5$	$C_1 = 14.7, C_2 = 30.0$	0.9791	0.9742	0.9896
	$\mu_2 = 0.8, \sigma_2 = 0.5$	$C_1 = 33.1, C_2 = 59.1$	0.9647	0.9647	0.9769

Table 7
Power simulation results for lognormal distribution at significance level $\alpha = 0.05$, from 10,000 replications with equal censoring.

Relative skewness	Survival time	Average % censored	Adaptive	Logrank	Wilcoxon
$n_1 = n_2 = 20$					
		$C_1 = 0.0, C_2 = 0.0$	0.3522	0.3189	0.3384
Heavy	$\mu_1 = 0, \sigma_1 = 1.5$	$C_1 = 16.3, C_2 = 27.9$	0.3040	0.2992	0.3157
	$\mu_2 = 0.75, \sigma_2 = 1.5$	$C_1 = 31.3, C_2 = 47.1$	0.2755	0.2720	0.2891
		$C_1 = 0.0, C_2 = 0.0$	0.6211	0.5743	0.6207
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 11.1, C_2 = 21.9$	0.5765	0.5556	0.5919
	$\mu_2 = 0.75, \sigma_2 = 1$	$C_1 = 25.4, C_2 = 45.1$	0.5235	0.5161	0.5427
		$C_1 = 0.0, C_2 = 0.0$	0.8470	0.8146	0.8539
Light	$\mu_1 = 0, \sigma_1 = 0.75$	$C_1 = 8.8, C_2 = 18.5$	0.8150	0.7931	0.8282
	$\mu_2 = 0.75, \sigma_2 = 0.75$	$C_1 = 21.7, C_2 = 42.8$	0.7634	0.7520	0.7849
$n_1 = n_2 = 50$					
		$C_1 = 0.0, C_2 = 0.0$	0.3805	0.3445	0.3721
Heavy	$\mu_1 = 0, \sigma_1 = 1.5$	$C_1 = 25.7, C_2 = 35.5$	0.3235	0.3204	0.3407
	$\mu_2 = 0.5, \sigma_2 = 1.5$	$C_1 = 39.4, C_2 = 50.7$	0.2955	0.2951	0.3122
		$C_1 = 0.0, C_2 = 0.0$	0.6790	0.6237	0.6799
Moderate	$\mu_1 = 0, \sigma_1 = 1$	$C_1 = 19.7, C_2 = 30.1$	0.6041	0.5878	0.6343
	$\mu_2 = 0.5, \sigma_2 = 1$	$C_1 = 35.0, C_2 = 49.8$	0.5570	0.5537	0.5902
		$C_1 = 0.0, C_2 = 0.0$	0.8895	0.8508	0.8954
Light	$\mu_1 = 0, \sigma_1 = 0.75$	$C_1 = 16.5, C_2 = 26.7$	0.8457	0.8176	0.8654
	$\mu_2 = 0.5, \sigma_2 = 0.75$	$C_1 = 31.9, C_2 = 48.2$	0.7860	0.7749	0.8243

the Wilcoxon is better than the logrank. This is probably because the hazard function of the lognormal distribution is very similar to that of the loglogistic distribution (Klein and Moeschberger, 2003).

The adaptive test tends to fall somewhere in the middle under all distributions considered. The adaptive test is better than logrank under loglogistic and lognormal distributions, and better than Wilcoxon under the Weibull.

5. Application to catheter placement study

Let us consider the data on a clinical trial of the effectiveness of two catheterization procedures in kidney dialysis patients, taken from page 6 of Klein and Moeschberger (2003). The time to first exit-site infection (in months) was observed from 43

patients who utilized a surgically placed catheter (Group 1) and 76 patients who utilized a percutaneous placed catheter (Group 2).

We shall perform the adaptive survival test using the R code in the Appendix to test if there is a difference in the time to infection between the two groups. The calling program standardizes the two samples and calls the *getmle* function to obtain the $\hat{m}$ discrimination statistic. Based on $\hat{m}$, the *adaptivetest* function selects between logrank and Peto–Peto's Wilcoxon test statistics, which are both derived from Harrington and Fleming's G^{ρ} family of tests (1982). The likelihood function is maximized at $\hat{m} = 50\,359.12$, which suggests that Weibull distribution is a better fit to the data than the loglogistic distribution. Therefore, the logrank test ($z = 1.59$, $p = 0.112$) is chosen over the Wilcoxon test ($z = 1.18$, $p = 0.239$).

6. Conclusion

The simulation results have shown that the adaptive two-sample test is more robust to the underlying survival distribution, relative to the logrank and Wilcoxon tests. Under various distribution and censoring configurations, it performed better than the less efficient of the logrank and Wilcoxon tests, while comparing favorably with the better one. It seems to do this without substantial inflation of Type I error. We believe the discriminant diagnostic based on $\hat{m}$ can be a useful pretest for choosing between logrank and Wilcoxon tests, the same way a test for common variance is used to discriminate between the Welch two-sample t-test and the pooled-variance t-test.

The adaptive test can be extended to the K-sample problem by modifying the procedure's step which standardizes the failure times from the two samples. Such K-sample adaptive test can then be further developed into a stratified test to account for covariates.

This paper only investigated the performance of the adaptive test on three well-known survival distributions: Weibull, loglogistic and lognormal. Their performance on other survival distributions is being evaluated in subsequent research.

Appendix. R code for adaptive test

See http://www.stat.wmich.edu/naranjo/adaptiveRcode/.

```
## Log likelihood function to be maximized

fr <- function(theta,y){
mean<-theta[1]
sigma<-theta[2]
m<-theta[3]
v <- (log(y)-mean)/sigma
sampsize <- length(y)
logl<--sampsize*(log(m))+sum(v)-(1+m)*sum(log(1+(1/m)*exp(v)))-
      sampsize*log(sigma)-sampsize*log(beta(1,m))
return(logl)
}

## Gradient function

gr <- function(theta,y){
mean<-theta[1]
sigma<-theta[2]
m<-theta[3]
sampsize <- length(y)
v <- (log(y)-mean)/sigma
s <--log(m)+v
gradient<-numeric(length(theta))
gradient[1]<--(sampsize/sigma) + ((1+m)/sigma)*sum(exp(s)/(1+exp(s)))
gradient[2]<--(1/sigma)*sum(v) + ((1+m)/sigma)*sum(exp(s)*v/(1+exp(s)))-
             (sampsize/sigma)
gradient[3]<--(sampsize/m)-sampsize*(digamma(m)-digamma(1+m))+
             ((1+m)/m)*sum(exp(s)/(1+exp(s)))-sum(log(1+exp(s)))
return(gradient)
}
```

```
## Function that computes for the MLE of m, mean and sigma

getmle <- function(time){

xbar <- mean(log(time))
sd <- sd(log(time))
maxiter <-20000
maxxbar <- xbar+maxiter
maxsd <- sd+maxiter

for (initialmean in xbar:maxxbar){
     for (initialsigma in sd:maxsd){
          for (initialm in 1e-10:maxiter){
                mle<-try(optim(c(initialmean,initialsigma,initialm),
                      method="L-BFGS-B", fn=fr,gr=gr,
                      control=list(fnscale=-1,trace=1),
                      lower=c(-Inf,0,1e-10), upper=rep(Inf,3),
                      hessian=TRUE,y=time))

                if (class(mle) == "try-error") {next}
                else {if (mle$convergence != 0) {next}
                           else {break}
                      }
          }
          if (class(mle) != "try-error") {
             if (mle$convergence != 0) {next}
                else {break}
          }
     }
   if (class(mle) != "try-error") {
      if (mle$convergence != 0) {next}
         else {break}
   }
}

 list(meanhat=mle$par[1],sigmahat=mle$par[2],mhat=mle$par[3],
 estlogL = mle$value,convergence=mle$convergence)
}

## Function for the adaptive test

adaptivetest <- function(data,m){
LR <- survdiff(Surv(time,status==1)~group,data=data,rho=0)
WIL <- survdiff(Surv(time,status==1)~group,data=data,rho=1)
LRztest <- sqrt(LR$chisq)
Wilztest <- sqrt(WIL$chisq)
if (1/m > 0.5) newstat = Wilztest else newstat = LRztest
list(r=1/m,newstat=newstat,LRztest=LRztest,Wilztest=Wilztest)
}

--------------------------- CALLING PROGRAM-----------------------------
Let "combdata" be a data frame which consists of at least three vectors:
"time" (survival/failure time), "status" (0=censored,1=uncensored) and
"group" (1=Group 1,2=Group 2).

uncensoredT1 <- combdata$time[combdata$status==1 & combdata$group==1]
uncensoredT2 <- combdata$time[combdata$status==1 & combdata$group==2]
uncensoredT1_rescaled <- uncensoredT1*(sd(uncensoredT2)/sd(uncensoredT1))
mle <- getmle(time=c(uncensoredT1_rescaled,uncensoredT2))
output <- adaptivetest(data=combdata,m=mle$mhat)
```

References

Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer-Verlag, New York.

Beltangady, M.S., Frankowski, R.F., 1989. Effect of unequal censoring on the size and power of the logrank and Wilcoxon types of tests for survival data. Statistics in Medicine 8, 937–945.

Breslow, N., 1970. A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika 57, 579–594.

Cox, D.R., 1972. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187–220.

Davis, P.J., 1965. Gamma function and related functions. In: Abramowitz, M., Stegun, I.A. (Eds.), Handbook of Mathematical Functions. Dover Publications Inc., New York, pp. 253–266.

Gehan, E.A., 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika 52, 203–223.

Gill, R.D., 1980. Censoring and Stochastic Integrals. In: Mathematical Centre Tract, vol. 124. Mathematisch Centrum, Amsterdam.

Harrington, D.P., Fleming, T.R., 1982. A class of rank test procedures for censored survival data. Biometrika 69, 553–566.

Hess, K., 1995. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Statistics in Medicine 14, 1707–1723.

Kalbfleisch, J.D., Prentice, R.L., 1980. The Statistical Analysis of Failure Time Data. John Wiley and Sons Inc., New York.

Klein, J.P., Moeschberger, M.L., 2003. Survival Analysis: Techniques for Censored and Truncated Data. Springer, New York.

Latta, R., 1981. A Monte Carlo study of some two-sample rank tests with censored data. Journal of the American Statistical Association 76, 713–719.

Lee, E.T., Desu, M.M., Gehan, E.A., 1975. A Monte Carlo study of the power of some two-sample tests. Biometrika 62, 425–432.

Leton, E., Zuluaga, P., 2001. Equivalence between score and weighted tests for survival curves. Communications in Statistics-Theory and Methods 30, 591–608.

Leton, E., Zuluaga, P., 2005. Relationships among tests for censored data. Biometrical Journal 47, 377–387.

Mantel, N., 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports 50, 163–170.

Peto, R., 1972. Rank tests of maximal power against Lehmann-type alternatives. Biometrika 59, 472–475.

Peto, R., Peto, J., 1972. Asymptotically efficient rank invariant test procedures (with discussion). Journal of the Royal Statistical Society, Series A 135, 185–207.

Prentice, R.L., 1975. Discrimination among some parametric models. Biometrika 62, 607–614.

Prentice, R.L., 1978. Linear rank tests with right censored data. Biometrika 65, 167–179.

Prentice, R.L., Marek, P., 1979. A qualitative discrepancy between censored data rank tests. Biometrics 35, 861–867.