

The Weighted Log-Rank Class of Permutation Tests: P-values and Confidence Intervals Using Saddlepoint Methods

Ehab F. Abd-Elfattah
Ain Shams University, Cairo, Egypt

Ronald W. Butler
Colorado State University, Fort Collins, CO.

September 30, 2005

Abstract

Test statistics from the weighted log-rank class are commonly used to compare treatment with control when there is right censoring. This paper uses saddlepoint methods to determine mid- p -values from the null permutation distributions of tests from the weighted log-rank class. Analytical saddlepoint computations replace the permutation simulations and provide mid- p -values that are virtually exact for all practical purposes. The speed of these saddlepoint computations makes it practical to invert the weighted log-rank tests to determine nominal 95% confidence intervals for the treatment effect with right censored data. Such analytical inversions lead to permutation confidence intervals that are easily computed and virtually identical with the exact intervals that would normally require massive amounts of simulation.

Some key words: Linear rank test; log-rank test, mid- p -value; permutation distribution; saddlepoint approximation; symmetry test; weighted log-rank class.

1 Introduction

Linear rank statistics are often used to test the effectiveness of a treatment as compares to a control in the two independent samples context. In clinical trials, the time to event responses are typically right censored and modified tests from the class of weighted log-rank statistics are most often used to accommodate the

censoring. The most commonly used rank tests from this class are the log-rank and generalised Wilcoxon (Peto-Prentice) tests (Gehan, 1965; Peto and Peto, 1972; Prentice, 1978). The presentation below concerns the entire weighted log-rank class and uses as examples the log-rank and generalised Wilcoxon tests, as well as Gehan's (1965) original test and specific tests from the Tarone-Ware (1977) and Fleming-Harrington (1981) classes.

The current paper proposes the use of saddlepoint approximations as a means for determining the significance levels for tests in the weighted log-rank class under their exact permutation distributions. Permutation significance was originally advocated by Peto and Peto (1972) however current software such as SAS uses asymptotic normal approximations as described, for example, in Kalbfleisch and Prentice (2002). It will be seen that saddlepoint approximations are almost always closer to the true permutation significance levels than normal approximations. The degree of greater accuracy is readily apparent in smaller and intermediate size samples for which the asymptotic normality has not been attained. A variety of examples and simulations with various log-rank statistics are used to show that the saddlepoint mid- p -value is an extremely accurate approximation for the mid- p -value as determined from the exact permutation distribution.

The focus is on mid- p -values rather than ordinary p -values because a major aim of this paper is to construct confidence intervals for the treatment effect through the inversion of the tests. When mid- p -values are used instead of ordinary p -values in this inversion, the intervals that result have true coverages that tend to be much closer to the nominal coverage. This fact has been discussed extensively in Agresti (1992), Kim and Agresti (1995), and Butler (2005, chapter 6). If the mid- p -values of rank tests are inverted by using saddlepoint approximations, the resulting confidence intervals are almost identical to the exact intervals determined by the massive simulations needed for the exact permuta-

tion distributions. Confidence intervals from the inversion of normal tests are described in Kalbfleisch and Prentice (2002) and are consistently less accurate. A range of practical examples is considered and in each case the saddlepoint intervals almost exactly replicate the true intervals from the exact permutation distribution.

For settings that lack censoring, Garthwaite (1996) has attempted to invert randomisation tests by using simulation in conjunction with a Robbins-Munro search process to locate the two ends of the confidence interval. Also Tritchler (1984) inverted probability generating functions of some simple permutation distributions by using the fast Fourier transform.

The computational methods based on saddlepoint approximation are extremely stable and have been programmed as a general purpose “black box” procedure. The executable files with instructions for use are available at <http://www.stat.colostate.edu/~walrus/>. Mid- p -values are computed for all five of the weighted log-rank tests exemplified in the paper and confidence intervals can also be determined through the inversion of these five tests.

Section 2 provides an overview of the weighted log-rank tests along with the associated permutation distributions that determine their mid- p -values. Saddlepoint approximation to these permutation distributions is addressed in section 3. Section four provides numerical examples along with extensive simulations that demonstrate the extraordinary accuracy of the saddlepoint approximations. Section five considers test inversion for confidence intervals and provides many examples. Section six indicates the modifications needed for the treatment of ties. Section seven concludes by showing how these saddlepoint approximations may be used to implement permutation tests for symmetry in the presence of censoring.

2 Weighted Log-Rank Class

Suppose sample size N_1 for the treatment group and N_2 for control group with $N = N_1 + N_2$. The pooled data are $\{(t_i, z_i, \delta_i) : i = 1, \dots, N\}$ where t_i is a time to event, z_i is a treatment indicator, and δ_i indicates an observed failure time. Assume independent censoring with the censoring distribution not dependent on group membership. With the survival functions for treatment and control as $S_1(t)$ and $S_2(t)$ respectively, then a test for $H_0 : S_1(t) \equiv S_2(t) = S(t)$ versus a one- or two-sided alternative is generally based on tests from the weighted log-rank class.

Let $t_{(1)} < t_{(2)} < \dots < t_{(k)}$ be the distinct ordered failure times among the pooled data with t_{i1}, \dots, t_{im_i} as the right censored times in the intervals $[t_{(i)}, t_{(i+1)})$, $i = 0, 1, \dots, k$ where $t_{(0)} = -\infty$, $t_{(k+1)} = \infty$ and $k + \sum_{i=1}^k m_i = N$. Also, let $z_{(i)}$ and z_{ij} for $i = 1, \dots, k$; $j = 1, \dots, m_i$ represent the corresponding indicators of treatment group membership. Assuming no ties among the uncensored data from different groups, the general weighted log-rank class of statistics is written as

$$v = \sum_{i=1}^k w_i \left(z_{(i)} - \frac{1}{n_i} \sum_{l \in R(t_{(i)})} z_l \right) \quad (1)$$

where n_i is the total number of individuals at risk at time $t_{(i)}^-$, and $R(t_{(i)})$ is the set of individuals at risk at $t_{(i)}^-$.

In most applications the weight function w_i is a fixed function of the risk set sizes $\{n_1, n_2, \dots, n_i\}$ up to time $t_{(i)}$. Among such tests are the log-rank test ($w_i \equiv 1$) with optimal power against the proportional hazard alternative, Gehan's (1965) test ($w_i = n_i$), the Tarone and Ware (1977) class $\{w_i = f(n_i)\}$ with specific recommendation ($w_i = \sqrt{n_i}$) considered here, the weight function

$$w_i = \prod_{j=1}^i \frac{n_j}{n_j + 1}$$

suggested in Peto and Peto (1972) and Prentice (1978) and referred to as the

generalised Wilcoxon, and the general class of tests of Fleming and Harrington (1981) with weight function

$$w_i = \hat{S}(t_{(i-1)})^p \{1 - \hat{S}(t_{(i-1)})\}^q \quad p \geq 0, q \geq 0.$$

Survival estimate $\hat{S}(t_{(i)})$ is the Kaplan-Meier estimator at time $t_{(i)}$ and the specific example $p = 1$ and $q = 0$ is considered here.

In the randomisation used for the permutation distribution of v , the failure times and censored times remain fixed in time order while the N_1 treatment labels are randomly assigned to $\binom{N}{N_1}$ of these time positions. In order to simplify saddlepoint approximation for this permutation distribution, it is expedient to rewrite v in the linear form

$$v = \sum_{i=1}^k \left\{ c_i z_{(i)} + C_i \sum_{j=1}^{m_i} z_{ij} \right\} \quad (2)$$

where the constants c_i and C_i are fixed constants that depend only on their time position $t_{(i)}$.

Proposition 1 *The weighted log-rank statistic v in (1) has a null permutation distribution given as the distribution of (2) where $z_{(1)}, \{z_{ij}\}, \dots, z_{(k)}, \{z_{kj}\}$ are treatment group indicators with the uniform distribution $\binom{N}{N_1}^{-1}$ over values for which $\sum_{i=1}^k (z_{(i)} + \sum_{j=1}^{m_i} z_{ij}) = N_1$. The weights in (2) are*

$$c_i = w_i - \sum_{l=1}^i \frac{w_l}{n_l}, \quad C_i = - \sum_{l=1}^i \frac{w_l}{n_l}.$$

Proof.

$$\begin{aligned} v &= \sum_{i=1}^k w_i \left\{ z_{(i)} - \frac{1}{n_i} \sum_{l=i}^k \left(z_{(l)} + \sum_{j=1}^{m_l} z_{lj} \right) \right\} \\ &= \sum_{i=1}^k w_i z_{(i)} - \sum_{i=1}^k \left(\sum_{l=1}^i \frac{w_l}{n_l} \right) \left\{ z_{(i)} + \sum_{j=1}^{m_i} z_{ij} \right\} \\ &= \sum_{i=1}^k \left(c_i z_{(i)} + C_i \sum_{j=1}^{m_i} z_{ij} \right). \end{aligned}$$

3 Saddlepoint Approximation for the Permutation Distribution

A uniform distribution on the treatment indicators $\{z_{(i)}\} \cup \{z_{ij}\}$ may be constructed from a corresponding set of independent and identically distributed Bernoulli (θ) indicator variables denoted in capitals as $\{Z_{(i)}\} \cup \{Z_{ij}\}$. Define Y to be the same weighting as in (2) and let X be the total count so that

$$Y = \sum_{i=1}^k \left\{ c_i Z_{(i)} + C_i \sum_{j=1}^{m_i} Z_{ij} \right\}$$

$$X = \sum_{i=1}^k \left\{ Z_{(i)} + \sum_{j=1}^{m_i} Z_{ij} \right\}.$$

For any $\theta \in (0, 1)$, the conditional distribution of Y given $X = N_1$ is the required permutation distribution which can be approximated by using the double saddlepoint approximation of Skovgaard (1987).

Previously single saddlepoint approximations for permutation distributions have been suggested by Daniels (1955), Robinson (1982), and Davison and Hinkley (1988). Double saddlepoint approximations for conditional distributions of the type given above were suggested by Daniels (1958) and further developed in Booth and Butler (1990).

Let P be a random variable with the required permutation distribution and v_0 the observed value of v . The mid- p -value is $\text{pr}(P > v_0) + \text{pr}(P = v_0)/2 = \text{mid-}p(v_0)$ and is approximated from the Skovgaard (1987) saddlepoint procedure as the conditional tail probability $\text{pr}(Y > v_0 | X = N_1)$. This approximation uses the joint cumulant generating function for (X, Y) given by $K(s, t) = \log M_{X,Y}(s, t)$ with

$$M_{X,Y}(s, t) = \prod_{i=1}^k [\{1 - \theta + \theta \exp(s + c_i t)\} \{1 - \theta + \theta \exp(s + C_i t)\}^{m_i}]. \quad (3)$$

Then

$$\text{mid-}p(v_0) = \Pr(Y > v_0 | X = N_1) \simeq 1 - \Phi(\hat{w}) - \phi(\hat{w}) \left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}} \right) \quad (4)$$

where

$$\begin{aligned} \hat{w} &= \text{sgn}(\hat{t}) \sqrt{2 \left[\{K(\hat{s}_0, 0) - N_1 \hat{s}_0\} - \{K(\hat{s}, \hat{t}) - N_1 \hat{s} - v_0 \hat{t}\} \right]} \\ \hat{u} &= \hat{t} \sqrt{|K''(\hat{s}, \hat{t})| / K''_{ss}(\hat{s}_0, 0)}. \end{aligned}$$

In these expressions, K'' is the 2×2 Hessian matrix and K''_{ss} is the $\partial^2 / \partial s^2$ component of this Hessian. The numerator saddlepoint (\hat{s}, \hat{t}) solves

$$\begin{aligned} K'_s(\hat{s}, \hat{t}) &= \sum_{i=1}^k \left\{ \frac{\exp(\hat{s} + c_i \hat{t})}{(1 - \theta)/\theta + \exp(\hat{s} + c_i \hat{t})} + \frac{m_i \exp(\hat{s} + C_i \hat{t})}{(1 - \theta)/\theta + \exp(\hat{s} + C_i \hat{t})} \right\} = N_1 \\ K'_t(\hat{s}, \hat{t}) &= \sum_{i=1}^k \left\{ \frac{c_i \exp(\hat{s} + c_i \hat{t})}{(1 - \theta)/\theta + \exp(\hat{s} + c_i \hat{t})} + \frac{m_i C_i \exp(\hat{s} + C_i \hat{t})}{(1 - \theta)/\theta + \exp(\hat{s} + C_i \hat{t})} \right\} = v_0 \end{aligned}$$

while the denominator saddlepoint \hat{s}_0 solves

$$K'_s(\hat{s}_0, 0) = \frac{N \exp(\hat{s}_0)}{(1 - \theta)/\theta + \exp(\hat{s}_0)} = N_1. \quad (5)$$

Since the computations of \hat{w} and \hat{u} do not depend on the particular value of θ used, the value $\theta = N_1/N$ has been used since it results in an explicit solution for (5) as $\hat{s}_0 = 0$ and simplifies the calculations. For further discussion about this approximation, see Butler (2005).

The saddlepoint expression in (4) uses the saddlepoint approximation as if Y , and consequently P , were continuous random variables. In the permutation setting however, P is discrete and not even a lattice distribution for which a continuity correction would be available. The reason that this continuous formula can and should be used is that it provides the most accurate approximation for the mid- p -value. Pierce and Peters (1992), Davison and Wang (2002), and Butler (2005, §6.1.4) discuss reasons for this accuracy. Perhaps the simplest explanation in the last reference takes the view that a continuous saddlepoint

approximation is an approximation to the true inverse Fourier transform that determines $\text{pr}(P > v)$. Since $\text{pr}(P > v)$ has a step discontinuity at v_0 , the exact Fourier inversion at $v = v_0$ is the midpoint of the step (see Theorem 10.7b, Henrici, 1977) or the mid- p -value which is what the continuous saddlepoint approximation is actually approximating.

Agresti (1992), Routledge (1994) and Kim and Agresti (1995) have advocated use of the mid- p -value over the p -value since the ordinary p -value is too conservative. This claim finds its strongest justification when the significance tests are inverted to provide confidence intervals for the parameters under test. The use of 2.5% mid- p -values in either tail leads to confidence intervals whose nominal coverage is 95% and whose attained coverage is extremely close to this nominal coverage. This is not the case when the p -value is inverted and the attained coverage is consistently larger than the intended nominal coverage particularly with smaller sample sizes.

Since the determination of confidence intervals with the correct coverage in §5 is one of our main goals, the presentation focusses on mid- p -values to construct intervals with more accurate coverage. Also, if there is going to be any consistency in interpretation between significance levels and coverage probabilities, then the mid- p -value needs to be used in significance computation.

4 Numerical Examples and Simulations

Two published data sets are used to show the accuracy of the saddlepoint methods as compares to normal approximation in Table 1. The smaller data Set 1 is given in Kalbfleisch and Prentice (2002, p. 222) while the larger data Set 2 was used by Pike (1966) and Prentice (1978) and is reproduced in Kalbfleisch and Prentice (2002, p. 2). Table 1 summarizes the computation of the true (simulated) mid- p -values, saddlepoint mid- p -values, and normal p -values (which are

naturally mid- p -value approximations) for the five test statistics listed.

Data	N_1	N_2	Test stat.	True ¹ mid- p	Sadpt. mid- p	Normal p
All items treated as uncensored						
Set 1	4	5	LR	.01979	.02159	.01590
			GW	.02388	.02327	.02213
Set 2	21	19	LR	.03462	.03458	.03064
			GW	.04745	.04802	.04637
With Censored Items						
Set 1	4	5	LR	.01580	.01725	.00937
Censored:	1	1	GW	.01581	.01356	.01164
			GH	.01592	.01478	.01784
			TW	.01588	.01496	.01296
			FH	.01593	.01331	.01287
Set 2	21	19	LR	.05686	.05636	.04875
Censored:	2	2	GW	.05242	.05226	.04960
			GH	.06037	.06044	.06835
			TW	.05425	.05412	.06438
			FH	.05210	.05243	.05895

Table 1. True, saddlepoint, and normal mid- p -values for the log-rank (LR), generalised Wilcoxon (GW), Gehan (GH), Tarone-Ware (TW) and Fleming-Harrington (FH) statistics applied to the two sets of data. ¹Based on 10^6 simple random samples of N_1 from N and holding the censoring orders fixed.

The top four rows treat the censored values as actual survival times so that no censored values need to be accounted for. In each instance, the true mid- p -value has been calculated by taking 10^6 simple random samples of N_1 from N , holding the censoring orders fixed, and computing the proportion of times that P exceeds v_0 plus half the proportion of time it attains v_0 . With larger data sets the distinction between mid- p -value and p -value becomes negligible since the mass at v_0 is quite small. The saddlepoint approximation is highly accurate for both the smaller and larger data sets and also with and without censoring. By contrast, the normal approximation only works well with no censoring and shows inaccuracy with censoring even for the larger data Set 2.

4.1 Simulation Study

The saddlepoint accuracy seen in Table 1 occurs consistently over a wide range of conditions. For the log-rank and generalised Wilcoxon tests, simulation studies were conducted to evaluate the consistency of performance of saddlepoint mid- p -value approximation over a range of sample sizes, degrees of sample imbalance, and prevalence of censoring. Three error distributions were used to simulate data: log-logistic and Weibull for which the modified Wilcoxon and log-rank tests are the respective locally optimal tests, and a log-Weibull distribution. For each distribution, 1000 data sets were generated in the following way. First $N = N_1 + N_2$ independent and identically distributed responses were drawn from the distribution. Secondly, N_1 of these values were selected at random to determine the locations for treatment. Thirdly, the treatment values were translated to the right an amount $\beta > 0$ designed to induce borderline significance of a treatment effect, since this really is the most interesting situation. Finally a randomly chosen preset number of observations from each group were relabelled as censored. Small, intermediate and large sample sizes were used which were either balanced or unbalanced among the groups. The censoring percentage also changed between light, intermediate and heavy censoring. The aim in these choices was to keep the mean mid- p -value near .05.

Tables 2-4 provide summaries of these simulations for the three error distributions. Each table provides the following information: “Mean” is the average true mid- p -value based on 10^6 simulations for each of the 1000 data sets; “Sadpt. Prop.” is the proportion of the 1000 data sets for which the saddlepoint mid- p -value was closer to the true mid- p -value than the normal p -value; “Abs. Err. Sadpt.” is the average absolute error of the saddlepoint mid- p -value from the true

Stat.	Mean	Sadpt. Prop.	Abs. Err. Sadpt.	Abs. Err. Normal	Rel. Abs. Err. Sadpt.	Rel. Abs. Err. Nor.
$N_1 = 18$ $N_2 = 17$ $\beta = 1.5$ 30% censoring						
LR	.145	.983	.0 ³ 484	.0128	.0 ⁴ 100	.0 ³ 167
GW	.108	.980	.0 ³ 198	.0 ² 436	.0 ⁵ 300	.0 ³ 114
$N_1 = 18$ $N_2 = 17$ $\beta = 1.5$ 5% censoring						
LR	.059	.996	.0 ³ 413	.0 ² 353	.0 ⁵ 800	.0 ³ 115
GW	.041	.855	.0 ³ 132	.0 ³ 810	.0 ⁵ 100	.0 ⁴ 160
$N_1 = 8$ $N_2 = 7$ $\beta = 2.0$ 15% censoring						
LR	.117	.995	.0 ² 137	.0237	.0 ⁴ 140	.0 ³ 459
GW	.090	.982	.0 ³ 403	.0 ² 663	.0 ⁵ 830	.0 ⁴ 630
$N_1 = 36$ $N_2 = 34$ $\beta = 1.0$ 5% censoring						
LR	.122	.986	.0 ³ 343	.0 ² 509	.0 ⁶ 100	.0 ⁴ 130
GW	.087	.909	.0 ³ 178	.0 ² 111	.0 ⁵ 444	.0 ⁴ 300

Table 2. Performance under simulation from the log-logistic distribution.

Stat.	Mean	Sadpt. Prop.	Abs. Err. Sadpt.	Abs. Err. Normal	Rel. Abs. Err. Sadpt.	Rel. Abs. Err. Nor.
$N_1 = 12$ $N_2 = 8$ $\beta = 0.8$ 30% censoring						
LR	.085	.999	.0 ² 110	.0339	.0 ⁴ 254	.0 ³ 284
GW	.046	.997	.0 ³ 296	.0108	.0 ⁶ 128	.0 ³ 580
$N_1 = 18$ $N_2 = 17$ $\beta = 0.5$ 30% censoring						
LR	.054	.987	.0 ³ 376	.0 ² 502	.0 ⁵ 219	.0 ³ 221
GW	.028	.872	.0 ⁴ 886	.0 ³ 980	.0 ⁵ 794	.0 ³ 145
$N_1 = 36$ $N_2 = 34$ $\beta = 0.3$ 30% censoring						
LR	.09	.987	.0 ³ 221	.0 ² 387	.0 ⁶ 266	.0 ⁵ 273
GW	.06	.914	.0 ³ 150	.0 ² 110	.0 ⁵ 317	.0 ⁴ 377
$N_1 = 30$ $N_2 = 10$ $\beta = 0.55$ 30% censoring						
LR	.055	1.0	.0 ³ 303	.0186	.0 ⁴ 220	.0 ³ 472
GW	.032	1.0	.0 ³ 106	.0 ² 740	.0 ⁵ 400	.0 ³ 486

Table 3. Performance under simulation from the log-Weibull distribution.

mid- p -value; “Rel. Abs. Err. Sadpt.” is the average relative absolute error

of the saddlepoint mid- p -value from the true mid- p -value; and the remaining listings are the same assessments for the normal approximation.

Consider for example the simulation of 1000 data sets with $N_1 = 18$ and $N_2 = 17$ and 30% censoring. With the log-rank test, the saddlepoint mid- p -value was closer to the true value 98.3% of the time. For the generalised Wilcoxon test the saddlepoint approximation was closer 98% of the time. For the log-rank test, the absolute error of the saddlepoint approximation was 0.0484% versus 1.28% for the normal approximation, and the relative error the saddlepoint approximation was 0.001% versus 0.0167% for the normal approximation.

Stat.	Mean	Sadpt. Prop.	Abs. Err. Sadpt.	Abs. Err. Normal	Rel. Abs. Err. Sadpt.	Rel. Abs. Err. Nor.
$N_1 = 8$		$N_2 = 7$	$\beta = 1.5$	15% censoring		
LR	.049	1.0	.0 ³ 895	.0122	.0 ⁵ 100	.0 ³ 389
GW	.055	.83	.0 ³ 764	.0 ² 226	.0 ⁵ 900	.0 ⁴ 290
$N_1 = 18$		$N_2 = 17$	$\beta = 1.0$	15% censoring		
LR	.031	1.0	.0 ³ 247	.0 ² 435	.0 ⁴ 230	.0 ³ 607
GW	.044	.860	.0 ³ 121	.0 ³ 793	.0 ⁴ 230	.0 ⁴ 220
$N_1 = 32$		$N_2 = 38$	$\beta = 0.7$	30% censoring		
LR	.040	.979	.0 ³ 128	.0 ² 209	.0 ⁴ 160	.0 ³ 267
GW	.055	.856	.0 ³ 132	.0 ³ 746	.0 ⁴ 110	.0 ⁴ 650
$N_1 = 10$		$N_2 = 30$	$\beta = 1.0$	30% censoring		
LR	.048	.990	.0 ³ 212	.0 ² 646	.0 ⁴ 180	.0 ³ 490
GW	.064	.979	.0 ³ 147	.0 ² 403	.0 ⁵ 400	.0 ³ 223

Table 4. Performance under simulation from the Weibull distribution.

Overall, the saddlepoint approximation performed better than the normal approximation in all cases and the discrepancy was greater for the log-rank than the generalised Wilcoxon test. Inferior normal approximation to the permutation distribution of the log-rank test as compares with the generalised Wilcoxon test has also been noted by Heller and Venkatraman (1996). When averaged over all simulations, the saddlepoint approximation was closer 99.18% and 91.95%

of the time respectively for the log-rank and generalised Wilcoxon tests. In most cases the saddlepoint approximation demonstrated a relative error that is less than 0.001% and an absolute deviation of 0.05%. Unbalanced data, heavy censoring, and nonsymmetric error distributions had a much greater detrimental effect on the accuracy of the normal approximation than the saddlepoint approximation.

5 Confidence Interval for the Treatment Effect

Treatment effects are defined as locational shifts on the log-scale for the uncensored data model. If T is an uncensored failure time, then $\log T$ is assumed to have location parameters μ and $\mu + \beta$ respectively for the control and treatment groups that share a common error distribution. Let the (unordered) log-survival/censoring times be denoted using the N -vector $y = (\log t_1, \dots, \log t_N)^T$ with $z = (z_1, \dots, z_N)^T$ indicating the corresponding treatment group membership. The framework of the censored accelerated failure time model, as described in Kalbfleisch and Prentice (2002), determines the confidence interval for β . While the rank tests of §3 were concerned with testing $H_0 : \beta = 0$ essentially using the components of y , these same tests provide for testing $H_0 : \beta = \beta_0 \neq 0$ if the log-survival/censoring time residuals $y - z\beta_0$ are used in place of y . Within this framework, a 95% confidence interval consists of those β_0 values whose mid- p -values in (4) fall within the range [.025, .975].

Prentice (1978) has inverted such tests by using the asymptotic normal distribution theory for the standardized rank tests. Using a fine grid of β_0 -values with increment 0.01, he computed normal p -values for the log-rank and generalised Wilcoxon tests. As a function of increasing values of β_0 , the normalised value of v is a step function that makes an incremental decrease whenever the residual for a treatment subject is interchanged with the value of a control sub-

ject. When a 0.01 increment in the value of β_0 does not lead to an interchange, then the normal statistic and p -value remain unchanged and the dependence is flat.

The idea of inverting rank tests is conceptually simple and easy to implement, however there are some subtleties that need to be noted. First consider the determination of the required cutoff v_0 and, for purposes of discussion, suppose that $\beta_0 > 0$. For the treatment group all log-survival and log-censored times are diminished by amount β_0 which changes the relative ordering of treatment and control responses as well as the ordering of the positions held by censored observations. Denote the determination of the observed test statistic with treatment translation β_0 as $v_0(\beta_0)$. As previously mentioned with the normal approximation, the cutoff $v_0(\cdot)$ is a step function in β_0 that makes incremental decreases with increasing β_0 . Finally consider the permutation distribution of P whose distribution determines the mid- p -value as in (4). As noted in Proposition 1, the weights $\{c_i\}$ and $\{C_i\}$, used on ordered treatment survivals and censorings, depend on the relative ordering of the observed censored values from one group with the observed failure values of the other group. Thus when treatment values are shifted by amount $\beta_0 \neq 0$, $c_i = c_i(\beta_0)$ and $C_i = C_i(\beta_0)$ depend on the degree of shift β_0 . These weights determine the distribution of P so it also changes with β_0 , and is denoted as $P(\beta_0)$.

Under a β_0 translation of the treatment group, the mid- p -value is

$$p(\beta_0) = \text{pr}\{P(\beta_0) > v_0(\beta_0)\} + \frac{1}{2} \text{pr}\{P(\beta_0) = v_0(\beta_0)\}. \quad (6)$$

The saddlepoint approximation $\hat{p}(\beta_0)$ for (6) uses cutoff $v_0(\beta_0)$ in conjunction with the moment generating function in (3) for (X, Y) whose constants c_i and C_i are now replaced with $c_i(\beta_0)$ and $C_i(\beta_0)$. The set $\{\beta_0 : 0.025 \leq \hat{p}(\beta_0) \leq 0.975\}$ determines a 95% nominal saddlepoint interval.

5.1 Numerical Examples

The vaginal cancer data of Pike (1966) provides an example with an intermediate amount of data and light censoring. Take $y_i = \log(t_i - 100)$ and use a grid of β_0 -values over $(-5, 5)$ with incremental step 0.001. For each β_0 value, the residuals $y - \beta_0 z$ were computed, ordered, and the normal and saddlepoint mid- p -values were computed for both the log-rank and generalised Wilcoxon statistics. For the smaller values of β_0 used to determine the left edge of the confidence interval, Figures 1 and 2 plot $\hat{p}(\beta_0)$ versus β_0 for the log-rank and generalised Wilcoxon statistics respectively. The figures provide the saddlepoint (dotted) and normal (solid) mid- p -values for the collection of one-sided tests of $H_0 : \beta = \beta_0$ versus $H_1 : \beta > \beta_0$ over the range $\beta_0 \in (-0.1, 0.055)$ with $\hat{p}(\beta_0) \in (0.0, 0.1)$.

The horizontal dashed line indicates a height of 0.025 and selects the left edge of the saddlepoint (normal) interval as -0.041 (-0.038) where it crosses the dotted (solid) step function. The two figures show that $\hat{p}(\beta_0)$, the saddlepoint mid- p -value in the one-sided test of $H_0 : \beta = \beta_0$ versus $H_1 : \beta > \beta_0$, is consistently larger than the corresponding asymptotic normal p -value for the same hypothesis.

Figures 3 and 4 plot $1 - \hat{p}(\beta_0)$ versus β_0 for the log-rank and generalised Wilcoxon test respectively. The saddlepoint mid- p -values in these figures are for the one-sided tests of $H_0 : \beta = \beta_0$ versus $H_1 : \beta < \beta_0$. For these plots, the saddlepoint and normal approximations are closer and the right edges of the resulting confidence intervals are also closer. Table 6 summarizes the confidence intervals that result for the Pike data.

Interval	Lower			Upper		
	True	Sadpt.	Normal	True	Sadpt.	Normal
Log-rank	-.0403	-.041	-.038	.4269	.427	.424
G. Wilcoxon	-.0369	-.038	-.025	.4139	.414	.414

Table 6. Confidence intervals for the Pike data.

The plots in Figures 1 and 2 that provide the left edge of the interval are more interesting. Here an exclusion of $\beta = 0$ to the left of the confidence interval would demonstrate that treatment is significant at mid- p -value 2.5% in the one-sided hypothesis test of $H_0 : \beta = 0$ versus $H_1 : \beta > 0$. These same plots are also the plots in which the saddlepoint determination of mid- p -value is more shifted away from the asymptotic normal determination. For the Pike data, the true mid- p -value is .05242, the saddlepoint mid- p -value is .05226 and the normal p -value is shifted away at .0496.

Figures 3 and 4 plot mid- p -values $1 - \hat{p}(\beta_0)$ of one-sided tests whose alternatives state that the β_0 -translated treatment responses fare worse than control. Here the saddlepoint and normal approximations are closer together, but this also is the less interesting tail for determining a beneficial treatment effect.

The true 95% confidence intervals were computed by using the simulated mid- p -values as described in §4. Since it was prohibitive to simulate mid- p -values over the entire grid of β_0 values, only the true mid- p -values for a sequence of β_0 values that searched for a root to $p(\beta_0) = 0.025$ were used. Starting with the saddlepoint confidence interval, this sequence of β_0 -values consisted of those needed when using a bisection method to solve for the root of $p(\beta_0) = 0.025$.

The three methods of confidence interval construction in Table 6 are all conservative. For each, the determination of endpoints is based on the vertical steps in the plot cutting through the horizontal line at height 0.025 and the coverage must include the full probability mass at the endpoints. It is clear that the saddlepoint confidence intervals are more accurate than the normal ones for both the log-rank and generalised Wilcoxon tests.

Any attempt to simulate these confidence intervals, for the purpose of determining whether the saddlepoint interval achieves more accurate coverage, would be difficult to implement and also difficult to interpret because the true coverage cannot be set to 95%. As indicated above, even the true confidence interval

using simulation is conservative due to the extra mass at the endpoints of the confidence interval that pushes the total coverage over 95%. However the simulations in §4.1 suggest that if (a, b) is the true $100\alpha\%$ confidence interval with $\alpha \approx 0.95$, then $\hat{p}(a) \simeq p(a)$ and $\hat{p}(b) \simeq p(b)$ so that $\hat{p}(b) - \hat{p}(a) \simeq p(b) - p(a) = \alpha$ and the total coverage as well as overshoot and undershoot should all be very close to their true values.

5.2 Further Examples

Some additional examples are given in Table 7 for some published data sets using $y_i = \log(t_i)$ so time 0 is the baseline for measurement. In each, β is the differential effect of the second group. The saddlepoint confidence interval “Sadpt.” is extremely accurate in all instances of the log-rank and

Interval	Lower			Upper		
	True	Sadpt.	Normal	True	Sadpt.	Normal
Breast Cancer: Sedmak et al. (1989)				$N_1 = 36(20)$	$N_2 = 9(1)$	
Log-rank	-2.1477	-2.113	-2.069	-.1940	-.194	-.278
G. Wilcoxon	-1.9387	-1.940	-1.953	.03498	.0350	.0400
Ovarian Cancer: Edmunson et al. (1979)				$N_1 = 13(6)$	$N_2 = 13(8)$	
Log-rank	-.7901	-.808	-.676	3.0348	3.035	2.351
G. Wilcoxon	-.5586	-.559	-.527	2.9519	2.952	2.256
Myelomatosis: Peto et al. (1977)				$N_1 = 12(6)$	$N_2 = 13(2)$	
Log-rank	-5.0230	-4.722	-4.056	1.7739	1.774	1.774
G. Wilcoxon	-4.1829	-4.183	-3.335	2.1699	2.170	2.170
Gastric Carcinoma: Stablein et al. (1981)				$N_1 = 47(9)$	$N_2 = 48(8)$	
Log-rank	-.2575	-.258	-.229	.9099	.910	.885
G. Wilcoxon	.0685	.068	.081	1.0039	1.004	1.000

Table 7. Four published data sets with exact 95% confidence intervals “True”, saddlepoint determination “Sadpt.”, and normal intervals “Normal”. Group sample sizes and the number censored (in parentheses) are shown.

generalised Wilcoxon inversion. The examples show large and small data sets with heavy and light censoring that is balanced and unbalanced among the groups.

From a practical point of view, the 95% confidence interval for β is more meaningfully reported as a 95% confidence interval for $100(e^\beta - 1)\%$, the percentage increase in treatment survival time over control survival time under the accelerated failure time model. To understand this, let T_1 and T_2 be control and treatment survival times respectively with ε as the error distribution on the log-scale. The ratio of means (or medians for that matter) is

$$\frac{E(T_2)}{E(T_1)} = \frac{e^{\mu+\beta} E(e^\varepsilon)}{e^\mu E(e^\varepsilon)} = e^\beta$$

so that $100(e^\beta - 1)\%$ is the percentage increase in survival time due to the treatment effect. Ninety-five percent confidence intervals for this percentage are output by the software available at <http://www.stat.colostate.edu/~walrus/>.

6 Tied Failure Times

Suppose $t_{(1)} < \dots < t_{(k)}$ are the ordered failure times, and there are d_i failures at time $t_{(i)}$. Let $z_{(i)j}$ be the treatment indicator for individual j at time $t_{(i)}$ and suppose z_{i1}, \dots, z_{im_i} are treatment indicators for the censored data in $[t_{(i)}, t_{(i+1)})$.

The weighted log-rank statistic for the tied data is

$$v_t = \sum_{i=1}^k w_i \left(\sum_{j=1}^{d_i} z_{(i)j} - \frac{d_i}{n_i} \sum_{l \in R(t_{(i)})} z_l \right) \quad (7)$$

where n_i is the number of individuals at risk at time $t_{(i)}^-$, and $R(t_{(i)})$ is the risk set at time $t_{(i)}^-$. A simple exercise shows that v_t can be written in the form of (2) as

$$v_t = \sum_{i=1}^k \left[\left\{ w_i - \sum_{j=1}^i \frac{w_j d_j}{n_j} \right\} \sum_{j=1}^{d_i} z_{(i)j} + \left\{ - \sum_{j=1}^i \frac{w_j d_j}{n_j} \right\} \sum_{l=1}^{m_i} z_{il} \right]$$

with

$$c_i = w_i - \sum_{j=1}^i w_j d_j / n_j \quad C_i = - \sum_{j=1}^i w_j d_j / n_j.$$

Thus all tied failures at $t_{(i)}$ are given the same weight. With appropriate scores $\{c_i\}$ and $\{C_i\}$ and repeats of weights assigned to ties, the permutation distribution of v_t can be approximated by using the Skovgaard expression as in §3.

Tied failure times are found in the kidney data of Nahman et. al. (1992) with $N_1 = 43$ and $N_2 = 76$. Table 8 compares saddlepoint and normal mid- p -values with exact mid- p -values determined by simulating 10^6 permutations of v_t .

Mid- p -value	Log-rank	G. Wilcoxon	Gehan	TW	HF
True	.05098	.1136	.4883	.2574	.1144
Sadpt.	.05122	.1134	.4891	.2569	.1144
Normal	.05587	.1184	.4818	.2628	.1195

Table 8. Mid- p -values for five tests in the log-rank class in the presence of tied failure times.

7 Permutation Tests for Symmetry

Tests for symmetry of the distribution of uncensored log-survival times are proposed when the data are subject to right censoring. Such tests are relevant when using the generalised Wilcoxon test, whose motivation rests on the assumption of log-logistic errors, however no tests for symmetry have been found in the literature when there is right censoring. This section shows how such tests may be performed by using the two-sample tests of §3. If $\log T$ is an uncensored log-survival time that has been centered about 0, then the weighted log-rank class of statistics may be used to test that the distribution of $\log T$ is symmetry about zero.

Initially suppose there is no censoring and $y_i = \log t_i$ for $i = 1, \dots, N$ are the unordered log-survival times that form a random sample from continuous distribution G . For this setting, Tajuddin (1994) has suggested testing the symmetry of G about 0 by using the two-sample Wilcoxon test. He suggests pooling the data as $|y_1|, \dots, |y_N|$ and using $\{i : y_i > 0\}$ as the designated “treatment” group. For $Y \sim G$ assumed to be symmetric, the test is based on the idea that $-Y|Y < 0$ has the same distribution as $Y|Y > 0$ so that $|y_1|, \dots, |y_N|$ are a random sample from a common distribution. Tajuddin shows that such tests are at least as powerful as the test of McWilliams (1990) who in turn shows his test to be more powerful than tests by Butler (1969), Rothman and Woodroffe (1972), and Hill and Rao (1977) for selected alternatives in the asymmetric lambda distribution class.

Modification of this idea to accommodate right censoring leads to the weighted log-rank class of tests for symmetry. Suppose $\{y_i\}$ are the unordered log-survival/censoring times that have been centred, perhaps by subtracting the median value. Treatment labels are assigned to $\{i : y_i > 0\}$ and control labels to the remainder. Now all of the absolute treatment data are combined with only the absolute survival times from the control group. Censored data from control are eliminated for the reasons given below. This pooled data $|y_1|, \dots, |y_m|$ retains treatment/control labels as well as the censoring labels for the treatment group. Now any test in the weighted log-rank class may be used to check equality of the distributions for the treatment and control groups.

Censored control observations are removed because they are not informative about symmetry. To understand why, consider a censored survival time centred at, for example, -3 . Its centred survival time falls in the range $(-3, \infty)$ which becomes $[0, \infty)$ upon taking its absolute value. Thus a censored control value should be treated as censored at 0 and therefore it is not in $R(t_{(1)})$ and never at risk for any terms in the computation of v in (1). Thus, censored control values

are treated as if they are not in the data set and therefore not informative. Of course the same may be said if the smallest treatment value were censored; it is not in $R(t_{(1)})$ and its presence or removal has not effect on the computation of v .

Good and Gaskins (1980) provide uncensored data measuring the percentage of silica for $N = 22$ chondrites meteors. The data are to be tested for symmetry about the value 29 which leads to $N_1 = 11 = N_2$. Using the Wilcoxon signed rank test, the true mid- p -value is .36156, the saddlepoint mid- p -value is .36166, and the normal approximation yields .42914.

Dinse (1982) provides censored survival times for patients with non-Hodgkins lymphoma. The asymptomatic portion of the data are used and centered about the median which leads to $N_1 = 15 = N_2$. There are no censored values below the median but 13 out of the 15 values above the median are censored. Table 9 provides mid- p -values that contrast the accuracy of the saddlepoint approximation with the poor performance of the normal approximation.

	True	Sadpt.	Normal
Log-rank	.09225	.09033	.05051
Gehan	.05206	.05160	.04942

Table 9. Mid- p -values for testing the symmetry of survival data that are censored.

References

- [1] Agresti, A. (1992). A survey of exact inference for contingency tables. *Statist. Sci.*, **7**, 131-153.
- [2] Booth, J.G. and Butler, R.W. (1990). Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika* **77**, 787-796.
- [3] Box, G.E.P. and Anderson, S.L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J. R. Statist. Soc. B* **17**, 1-34.

- [4] Butler, C.C. (1969). A test for symmetry using the sample distribution function. *Ann. Math. Statist.* **40**, 2211-2214.
- [5] Butler, R.W. (2005). *Saddlepoint Approximations with Applications*. Under review, Cambridge University Press.
- [6] Cox, D.R. (1958). The regression analysis of binary sequences. *J. R. Statist. Soc. B* **20**, 215-242.
- [7] Daniels, H. E. (1955). In discussion of Box and Anderson (1955), 27-28.
- [8] Daniels, H. E. (1958). In discussion of Cox (1958), 236-238.
- [9] Davison, A.C. and Hinkley, D.V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* **75**, 417-431.
- [10] Davison, A.C. and Wang, S. (2002). Saddlepoint approximations as smoothers. *Biometrika* **89**, 933-938.
- [11] Dinse, G. E. (1982). Nonparametric estimation for partially complete time and type of failure data. *Biometrics*, **38**, 417-431.
- [12] Edmunson, J.H., Fleming, T.R., Decher, D.G., Malkasian, G.D., Jorgenson, E.O., Jeffries, J.A., Webb, M.J. and Kvols, L.K. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer Treatment Reports* **63**, 241-7.
- [13] Fleming, T. and Harrington, D. P. (1981). A class of hypothesis tests for one and two samples censored survival data. *Comm. Statist. A* **10**, 763-794.
- [14] Garthwaite, P.H. (1996). Confidence intervals from randomization tests. *Biometrics* **52**, 1387-1393.
- [15] Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203-223.
- [16] Good, I. J. and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75**, 42-56.
- [17] Heller, G. and Venkatraman, E.S. (1996). Resampling procedures to compare two survival distributions in the presence of right-censored data. *Biometrics* **52**, 1204-1213.
- [18] Henrici, P. (1977). *Applied and Computational Complex Analysis. Vol. 2. Special Functions-Integral Transforms-Asymptotics-Continued Fractions*. Wiley: New York.
- [19] Hill, D.L. and Rao, R.V. (1977). Tests of symmetry based on Cramér-von Mises statistics. *Biometrika* **64**, 489-494.

- [20] Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed. New York: Wiley.
- [21] Kim, D. and Agresti, A. (1995). Improved exact inference about conditional association in three-way contingency tables. *J. Amer. Statist. Assoc.* **90**, 632-639.
- [22] McWilliams, T.P. (1990). A distribution-free test for symmetry based on a runs test. *J. Amer. Statist. Assoc.* **85**, 1130-1133.
- [23] Nahman, N.S., Middendorf, D.F., Bay, W.H., McElligott, R., Powell, S., and Anderson, J. (1992). Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: clinical results in 78 patients. *Journal of the American Society of Nephrology* **3**, 103-107.
- [24] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *J. Roy. Statist. Soc. A* **135**, 185-206.
- [25] Peto, R., Pike, M.C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S.V., Mantal, N., McPherson, K., Peto, J., Smith, P.G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *British Journal of Cancer* **35**, 1-39.
- [26] Pierce, D.A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *J. R. Statist. Soc B* **54**, 701-737.
- [27] Pike, M. C. (1966). A method of analysis of certain class of experiments in carcinogenesis. *Biometrics* **22**, 142-161.
- [28] Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167-179.
- [29] Robinson, J. (1982). Saddlepoint approximations for permutation tests and confidence intervals. *J. R. Statist. Soc. B* **44**, 91-101.
- [30] Rothman, E.D. and Woodroffe, M. (1972). A Cramér-von Mises type statistic for testing symmetry. *Ann. Math. Statist.* **43**, 2035-2038.
- [31] Routledge, R.D. (1994). Practicing safe statistics with the mid- p . *Canadian J. Statist.* **22**, 103-110.
- [32] Sedmak, D.D., Meineke, T.A., Knechtges D.S., and Anderson, J. (1989). Prognostic significance of cytokeratin-positive breast cancer metastases. *Modern Pathology* **2**, 516-520.
- [33] Skovgaard, I.M. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Prob.* **24**, 875-887.

- [34] Stablein, D., Carter, W., and Novak, J. (1981). The analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials* **2**, 149-159.
- [35] Tarone, R. and Ware J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156-160.
- [36] Tajuddin, I. H. (1994). Distribution-free test for symmetry based on Wilcoxon two-sample test. *J. Appl. Statist.*, **21**, 409-416.
- [37] Tritchler, D. (1984). On inverting permutation tests. *J. Am. Statist. Assoc.* **79**, 200-207.

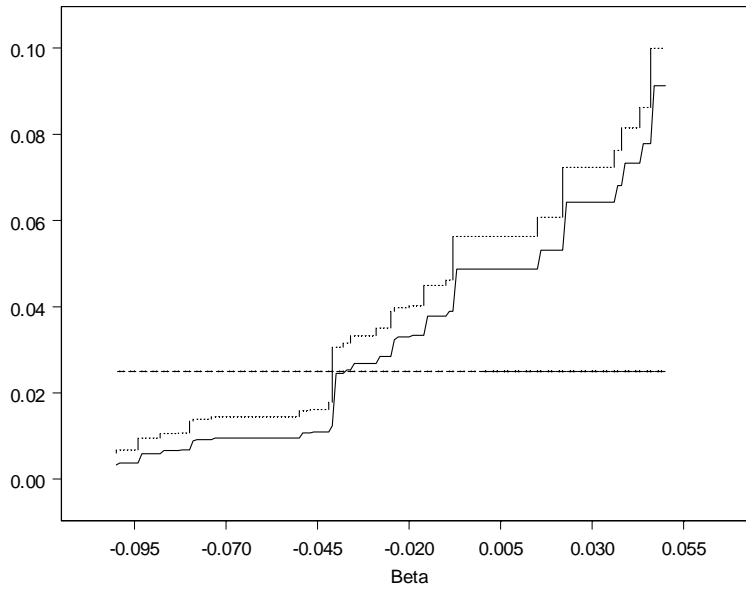


Figure 1. Plot of $\hat{p}(\beta_0)$ versus β_0 for the log-rank test using the Pike data. Mid- p -values for the saddlepoint approximations (dotted) and the normal approximations (solid) are shown over the range (0.0, 0.1).

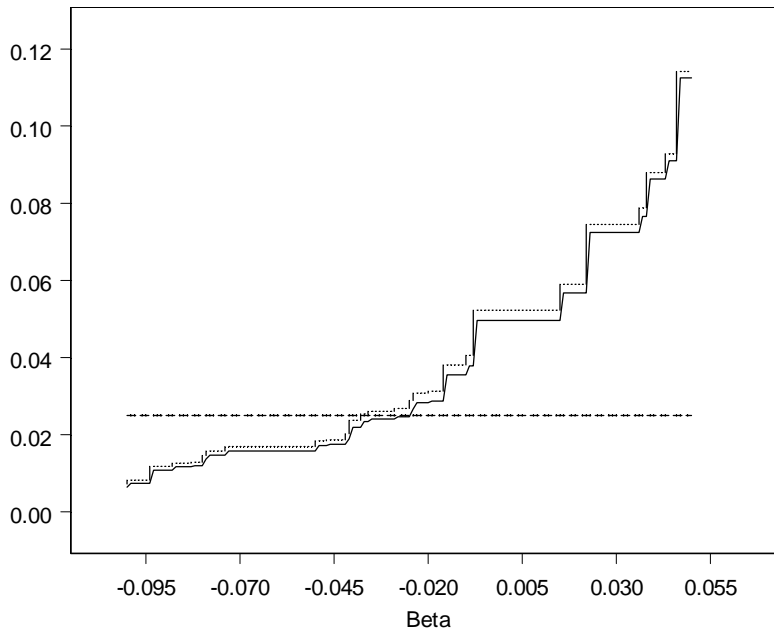


Figure 2. Same plot of $\hat{p}(\beta_0)$ versus β_0 as Figure 1 but for the generalized Wilcoxon test.

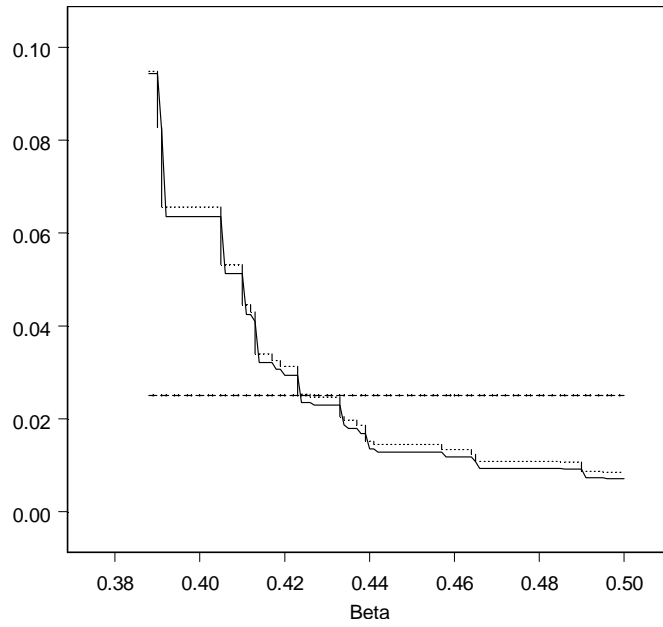


Figure 3. Plot of $1 - \hat{p}(\beta_0)$ versus β_0 for the log-rank test using the Pike data. Mid- p -values set against the alternative hypotheses $H_1 : \beta < \beta_0$ for the saddlepoint approximations (dotted) and the normal approximations (solid) are shown over the range $(0.0, 0.1)$.

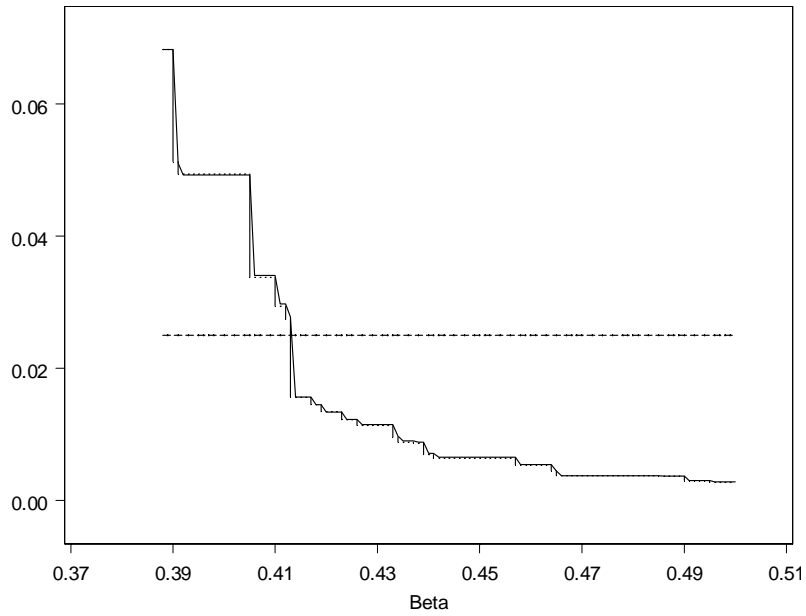


Figure 4. Same plot of $1 - \hat{p}(\beta_0)$ versus β_0 as Figure 3 but for the generalized Wilcoxon test.