

Rank invariant permutation tests and confidence intervals with interval-censored data

Ehab F. Abd-Elfattah

Department of Mathematics, Faculty of Education
Ain Shams University, Cairo, Egypt

Ronald W. Butler*

Department of Statistical Science, Southern Methodist University
Dallas, TX 75275, USA

February 24, 2010

Abstract

A large class of rank-invariant tests that includes the Peto and Peto (1972) class is considered for assessing the significance of a treatment vs. control in the presence of partially interval censored data. Saddlepoint approximations are used to approximate exact permutation mid- p -values for all tests in the rank-invariant class. The speed and stability of the saddlepoint computations make it practicable to invert these permutation tests to determine nominal $100(1-\alpha)\%$ confidence intervals for the treatment effect and the more clinically meaningful percentage increase in mean (median) treatment survival time as compares to control. Two other important ancillary results are also presented. First, an asymptotic sufficiency result is noted which says that asymptotically all discriminating information for assessing treatment significance rests in the permutation distribution of the treatment/control labels. Secondly, limitations of the hybrid ICM algorithm for determining NPMLEs of survival are noted and a modification is suggested to correct the problem.

Keywords: EM algorithm; ICM algorithm; Interval-censored; Mid- p -value; Permutation test; Rank-invariant; Saddlepoint approximation; Two sample tests.

1 Introduction

Interval censoring occurs in clinical trials and longitudinal studies when events of interest are assessed intermittently or at pre-scheduled times. In such situations,

*Corresponding Author: rbutler@mail.smu.edu, 214-768-1427

each event or survival time, is observed to occur within an interval of time. The special case of current status data, in which there is destructive testing or animal sacrifice during assessment, deals with a single assessment for an event of interest. In such cases the survival time has either occurred before the assessment time, in which case it is left-censored, or has not yet occurred, so it is right-censored. Both data types are considered as well as partially interval-censored data for which some exact survival times are observed. The data follow a two sample design in which a treatment group is compared with a control group with general (partial) interval censoring.

A large class of rank-invariant permutation tests is considered for assessing the significance of the treatment benefit. Mid- p -values for all of these tests are computed by using saddlepoint approximations that offer extremely accurate approximation to exact permutation significance levels in both small and large samples. For mid- p -values near 5%, an alternative approximation using simulation would require very large sample sizes to achieve comparable relative error. Normal approximations offer quite adequate approximation to significance levels of exact permutation tests in large samples. However, they are almost always less accurate than saddlepoint methods regardless of sample size.

Our methods extend the saddlepoint techniques, developed in Abd-Elfattah and Butler (2007) for the log-rank class of permutation tests dealing with right-censored data, to a general class of rank-invariant permutation tests proposed for use with general interval-censored data. Such tests include the Wilcoxon-type test proposed by Gehan (1965) and generalized by Mantel (1967), the general class of rank-invariant score tests proposed in Peto and Peto (1972), the score tests of Self and Grossman (1986) based on marginal likelihood, and a score test mimicking the ordinary log-rank test proposed by Finkelstein (1986) and Sun (1996). Additional tests are considered and can be found in Finkelstein (1986), Fay (1996), and Sun *et al.* (2005).

We also compare permutation significance with asymptotic normal significance, e.g. treatment significance determined by using the asymptotic normal distribution

of the test statistic under the null hypothesis. Early researchers were concerned with normal approximations dealing with the randomness associated with the permutation distribution of their test statistic with the aim of determining permutational significance. The more recent asymptotic theory for such test statistics in Sun *et al.* (2005) incorporates all the random variation of the censoring and survival processes into the limit theory. However, what we have discovered is that the same estimated asymptotic variances emerge for the normal limits in both instances, e.g. for permutation distribution approximation as well as for more general asymptotic limits. This suggests that asymptotically all discriminating information for assessing treatment significance rests in the permutation distribution associated with the treatment/control labels. Thus, regardless of intent, whether it be a determination of permutational or asymptotic normal significance, the same significance level results whenever a normal approximation is used.

Exact permutational significance becomes a more justifiable and meaningful computation than asymptotic normal significance when the assumptions underlying finite sample permutation methods are considered. While both require censoring that is not dependent on group labels, the asymptotic theory also requires independent censoring mechanisms as well as censoring mechanisms common to all subjects, assumptions not necessary for the permutation tests (Mantel, 1967). Our proposed saddlepoint methods focus on approximation to finite sample permutation significances thus they also avoid these additional assumptions.

The speed, accuracy, and stability of saddlepoint methods in determining permutational mid- p -values allow for the inversion of the interval-censoring tests to determine arbitrary level confidence intervals for treatment effect δ . Assume an accelerated failure time (AFT) model in which the treatment data on the log-scale are translated by treatment effect δ . Data subjected to such translation result in permutational significance $\hat{p}(\delta)$ which is plotted versus $\delta \in [-B, B]$ for some $B > 0$ so that a

100(1 - α)% confidence range

$$[\mathcal{L}, \mathcal{R}] = \{\delta : \alpha/2 \leq \hat{p}(\delta) \leq 1 - \alpha/2\} \quad (1)$$

can be determined for the treatment effect. The benefit of this is that the image of $[\mathcal{L}, \mathcal{R}]$ under the mapping $\delta \rightarrow 100(e^\delta - 1)$ represents a 100(1 - α)% confidence interval for the percentage increase in mean (or median) treatment survival time over control survival time in the AFT model. Such percentage increases provide more convincing evidence for the clinical importance of the treatment than significance levels alone.

These rank-invariant tests require computation of the nonparametric maximum likelihood estimate (NPMLE) for survival as discussed in Peto (1973) and Turnbull (1976). Indeed, confidence interval $[\mathcal{L}, \mathcal{R}]$ requires intensive use of such NPMLE computations since each $\hat{p}(\delta)$ requires a separate survival estimate $\hat{S}(\delta)$. Both the EM algorithm in Turnbull (1976) and the hybrid iterative convex minorant (hybrid ICM) in Wellner and Zhan (1997) failed to convergence to the NPMLE at some point during the course of our computations. To deal with this, we initially ran the EM algorithm and used its output as input for the hybrid ICM algorithm. By using both EM and ICM in this way, convergence to the NPMLE was always achieved in our computations.

General purpose programs that implement all methodology of the paper are available at <http://www.smu.edu/statistics/faculty/butler.html>. Executable files with instructions for use are provided to compute saddlepoint and normal approximations for permutation significance in the 7 tests considered. Additional programs also compute arbitrary level α confidence interval $[\mathcal{L}, \mathcal{R}]$ for the two most commonly used permutation tests.

The paper is organized as follows. Section 2 provides an historical review of the development of rank invariant tests for interval-censored data for the two sample model using sufficient detail that the 7 tests considered have explicit formulae. Permutational and asymptotic normal significance are compared in Section 3 and Section 4 briefly outlines how saddlepoint approximations are used to compute permutational

significance. Real data examples as well as simulations of permutational and normal significance computations are provided in Section 5 and test inversion for confidence interval determination is in Section 6. Section 7 concludes with discussion of NPMLE computations for survival, the failure of the Hybrid ICM algorithm and our solution for always finding NPMLEs.

2 Generalized rank-invariant tests

In a comparison of two groups, suppose a treatment group of n_1 is compared to a control group of n_2 with $n = n_1 + n_2$. Data from the pooled groups are $\{(l_i, r_i, z_i) : i = 1, \dots, n\}$ where $I_i = (l_i, r_i]$ is the range of time within which the i th survival is known to have occurred, and z_i is the indicator of treatment group membership. The model allows for the possibility of any combination of censoring and non-censoring including interval-censored observations ($l_i < r_i$), exactly observed survival times ($l_i = r_i^-$), and right- and left-censored observations ($r_i = \infty$ and $l_i = 0$ respectively). Let T_i denote the perhaps unobserved survival time for subject i .

If $S_1(t)$ and $S_2(t)$ are the respective survival functions for treatment and control groups, then rank-based permutation tests are considered for testing $H_0 : S_1(t) \equiv S_2(t) = S(t)$ versus a one-sided stochastically ordered alternative. The test statistics take the form

$$U = \sum_{i=1}^n z_i c_i, \tag{2}$$

where $\{c_i\}$ are various types of rank scores and reject H_0 in favour of $H_1 : S_1(t) > S_2(t)$ for small U . This one-sided permutational significance is determined by computing the left-tail probability for the observed value of $U = u$ using the permutation distribution of U , e.g. the uniform distribution on values of U under all $\binom{n}{n_1}$ distinct permutations of treatment labels $\{z_i\}$.

Gehan (1965) first proposes a Wilcoxon-type test in the class (2) when data are doubly censored and consist of left-censored, right-censored, or exactly observed values. Mantel (1967) extends the Gehan test to accommodate general interval censoring

and writes the test as (2) with score

$$c_i = \sum_{j=1, j \neq i}^n [1\{r_i \leq l_j\} - 1\{r_j \leq l_i\}] \quad (3)$$

that compares I_i to $\{I_j : I_j \cap I_i = \emptyset\}$ and counts the number of times I_i is below I_j minus the number of times I_i exceeds I_j .

Peto and Peto (1972, §4) formulate a large class of generalized rank-invariant tests that take the general form

$$U = U(\hat{S}) = \sum_{i=1}^n z_i \frac{\rho\{\hat{S}(l_i)\} - \rho\{\hat{S}(r_i)\}}{\hat{S}(l_i) - \hat{S}(r_i)}, \quad (4)$$

where the weight function ρ on $[0, 1]$ determines the specific test. Survival estimate \hat{S} is the NPMLE estimate of the survival function S under the null hypothesis computed by using the pooled set of interval-censored data $\{(l_i, r_i) : i = 1, \dots, n\}$ and making the assumption that $\hat{S}(\infty) = 0$ in (4). Also, for U to be meaningfully defined, only weight functions ρ are considered for which $\rho(0) = 0 = \rho(1)$ are defined by continuity.

Peto and Peto (1972, §4) derive this class of tests by assuming that $S_1 = S(\cdot; \theta_1)$ and $S_2 = S(\cdot; \theta_2)$ are members of a semi-parametric family of survival functions and by specifying H_0 as $\theta_1 = \theta_2 = \theta$. Taking the null survival $S = S(\cdot; \theta)$ as initially known and essentially an infinite dimensional nuisance parameter, they derive the locally most powerful score test of H_0 as based upon the statistic $U(S)$, where, in their examples, weight function ρ is determined by the parametric family and not dependent on S . The optimal rank invariant test results by plugging in the NPMLE estimate \hat{S} for unknown S and using test statistic $U = U(\hat{S})$. The rank invariance of statistic U is a consequence of the rank invariance of estimate \hat{S} .

The logistic family of distributions with θ as a location parameter generates the weight function $\rho(y) = y^2 - y$. In this context, Peto and Peto (1972, §6) derive a generalized Wilcoxon test so that the z_i weight in (4) is $\hat{S}(r_i) - \{1 - \hat{S}(l_i)\}$ and comparable to (3). Working with a general Lehmann family $\{S(t)^\theta : \theta > 0\}$ with S arbitrary, Peto and Peto (1972, §5.1) determine that $\rho(y) = y \ln y$ and construct what is now commonly known as the interval-censored log-rank test of H_0 . In their

rejoinder, Peto and Peto (1972) comment on the importance of this particular test in comparison to other possible tests. They emphasize that no other test can be locally more powerful than the permutational log-rank test against general Lehmann alternatives without sacrificing rank invariance.

When the data contain only right-censored and exact data, this interval-censored log-rank test turns out to not be the ordinary log-rank test developed for such data. It is however asymptotically equivalent to the latter (as $n \rightarrow \infty$) as was shown by Finkelstein (1986, §3).

Finkelstein (1986) also proposes the log-rank test in class (4) motivated from an equivalent point of view to Peto and Peto (1972). In the proportional hazards regression framework, with $\{z_i\}$ as an independent variable, Finkelstein (1986) shows that the score test for the significance of $\{z_i\}$, which profiles out the baseline survival by plugging in NPMLE \hat{S} , is her equation (8) which is exactly U with $\rho(y) = y \ln y$. This follows without any of the algebra in Finkelstein (1986) simply by the fact that her setup is the same as Peto and Peto (1972): her parametric family of distributions is the general Lehmann class with baseline survival playing the role as an infinite dimensional nuisance parameter profiled out by substituting \hat{S} .

Self and Grossman (1986) propose a rank invariant score test of the form (2) by working with an AFT model that treats $\{z_i\}$ as an independent variable. Following the marginal likelihood approach of Prentice (1978), their score test is computed from the marginal likelihood one would get under interval censoring, which is the probability of all ranks for the unobserved $\{\ln T_i\}$ that are consistent with their interval-censored $\{I_i\}$ values. Using a logistic error to generate rank scores leads to weights in (2) as

$$c_i = \sum_{k=m_i}^{M_i} w_i(k) \left(1 - \frac{2k}{n+1}\right) \quad (5)$$

where

$$m_i = 1 + \sum_{j=1}^n 1\{r_j \leq l_i\} \quad M_i = 1 + \sum_{j=1, j \neq i}^n 1\{l_j \leq r_i\} \quad (6)$$

define the minimum and maximum possible ranks¹ of $\ln T_i$ that are consistent with $\{I_i\}$, and $w_i(k)$ is the proportion of all possible rank vectors, consistent with $\{I_i\}$ for which $\ln T_i$ has rank k . Since the array of proportions $\{w_i(k)\}$ is very difficult to compute, Self and Grossman offer approximate alternatives such as their “Simple 2” option that is considered below in which

$$w_i(k) \propto \left(\sum_{i=1}^n 1\{m_i \leq k \leq M_i\} \right)^{-1},$$

or inversely proportional to the number of subjects whose range of possible ranks include rank k . When scores are generated by assuming errors with the extreme minimum value distribution, the weights for the test with the Simple 2 option are

$$c_i = \sum_{k=m_i}^{M_i} w_i(k) \left(1 - \sum_{j=1}^k \frac{1}{n-j+1} \right) \quad (7)$$

which are also considered below. See Kalbfleisch and Prentice (2002) equation 7.5 for the logistic weights in (5) and equation 7.4 for the extreme value weights in (7).

Finkelstein (1986, eqn. 12) and later Sun (1996) propose a test that makes use of imputations from the EM algorithm when it is used to compute NPMLE \hat{S} . The estimate \hat{S} only jumps in height on certain $(l, r]$ -bins where $l \in \{l_i\}$, $r \in \{r_i\}$, and $l < r$ are neighboring values from within the pooled set $\{l_i\} \cup \{r_i\}$; see Lindsey and Ryan (1998). Let $\{(\lambda_j, \rho_j] : j = 1, \dots, J\}$ be the $(l, r]$ -bins upon which \hat{S} places masses $\{\hat{\pi}_j = \hat{S}(\lambda_j) - \hat{S}(\rho_j)\}$. Then the test statistic is

$$U_{FS} = \sum_{j=1}^J (\Delta_{1j} - \Delta_j M_{1j}/M_j) \quad (8)$$

and mimics the standard log-rank test. Here, Δ_{1j} (Δ_j) is the imputed number of treatment (total) deaths in bin j from the EM algorithm, and M_{1j} (M_j) is the imputed number of treatment (total) subjects at risk in bin j imputed from the EM algorithm. To write (8) as a member of the class (2), denote $\alpha_{ij} = 1\{(\lambda_j, \rho_j] \subseteq (l_i, r_i]\}$, and let

$$\delta_{ij} = \frac{\alpha_{ij} \hat{\pi}_j}{\sum_{k=1}^J \alpha_{ik} \hat{\pi}_k}$$

¹In Self and Grossman (1986), $m_i = \vee r_i$ and $M_i = \wedge r_i$ but their description of these values differs from (6) and is incorrect.

be the imputed probability that subject i dies in bin j . Then

$$U_{FS} = \sum_{i=1}^n z_i \left\{ \sum_{j=1}^J (\delta_{ij} - \delta_{.j} m_{ij} / m_{.j}) \right\},$$

where $\delta_{.j} = \sum_{i=1}^n \delta_{ij} = \Delta_j$, $m_{ij} = \sum_{k=j}^J \delta_{ik}$, and $m_{.j} = \sum_{i=1}^n m_{ij} = M_j$ are (bin \times subject)-specific outputs from the EM algorithm. Finkelstein (1986) and Sun *et al.* (2005) comment that this test is asymptotically equivalent to the interval-censored log-rank test in class (4).

In the two sample context, Fay (1996) considers the linear transformation model

$$g(T_i) = \mu + \beta z_i + \varepsilon_i$$

with g as an increasing and smooth yet unspecified transformation and also leaving the distribution ε_i unspecified; see Kalbfleisch and Prentice (2002, §7.5). Fay treats g as a high dimensional nuisance parameter much as Finkelstein does with baseline survival function in her proportional hazards model, and as Peto and Peto do with their null survival function. His efficient scores test of $H_0 : \beta = 0$ leads to a test having the generalized rank-invariant test form in (4) with $\rho(y) = -(f \circ F^{-1})(1 - y)$, where ε_i has distribution F and density f . For logistic and extreme value error distributions, the respective weight functions are $y^2 - y$ and $y \ln y$ which leads to the same score tests recommended by Peto and Peto (1972).

Sun *et al.* (2005) also suggest the class of tests (4) and, motivated by the weights used by Fleming and Harrington (1991), extend the class to include tests of the form $\rho(y) = (y \ln y) y^\rho (1 - y)^\gamma$ for $\rho, \gamma \geq 0$. Perhaps the most important aspect of this paper is the derivation of useful asymptotic normal distributions under H_0 for statistics in the class (4) with asymptotic variances that can be reduced to simple expressions as given in the next section.

3 Permutational versus asymptotic normal significance

Permutation significances for tests in the class (2) are computed as tail probabilities for the observed value $U = u$ using the permutation distribution of U . This is the empirical distribution for U obtained by permuting the treatment indicators $\{z_i\}$ over all possible $\binom{n}{n_1}$ permutations while holding weights $\{c_i\}$ fixed. Advocates of this approach included all the early researchers such as Gehan (1965), Mantel (1967), and Peto and Peto (1972), as well as later researchers such as Self and Grossman (1986) and Fay (1996).

The validity of permutation significances is dependent on having “equal” censoring in each group. This can be expressed more precisely in terms of the joint distribution of the vector of (un)observables $\{L_i, R_i, T_i, Z_i\}$ associated with each subject. Even if there is random allocation of subjects to treatments, so $\{Z_i\}$ are exchangeable, this validity is not assured simply by assuming that the collection of four-vectors $\{L_i, R_i, T_i, Z_i\}$ forms an exchangeable sequence of n vectors. Under H_0 , what is needed is that $\{L_i, R_i, T_i\}$ are independent of $\{Z_i\}$ so that once allocated (given $Z_i = z_i$), the distribution of L_i, R_i, T_i does not depend on z_i . What is allowed is joint distributional dependence on the index i as well stochastic dependence between the censoring mechanism (L_i, R_i) and survival T_i . The former dependence allows for the possibility that individual subjects are censored according to their individual attributes and was discussed extensively by Mantel (1967). The latter dependence violates the traditional assumption of independent censoring mechanisms. As discussed in Mantel (1967), with sufficiently large samples, heterogeneity in the distributions of $\{L_i, R_i, T_i\}$ can be accounted for and will balance out in the treatments as a result of random allocation.

For early researchers, asymptotic normal approximations were a means for approximating these permutation distributions. Indeed, such proofs were based strictly

upon permutation randomization with the typical proof showing that all standardized moments of U converge to those of a standard normal; see Chung (1974) for a statement of the method of moments. An asymptotic approximation to the permutation variance for U is

$$v_p = \frac{1}{n-1} \left(\sum_{i=1}^n c_i^2 \right) \sum_{i=1}^n (z_i - \bar{z})^2 = \frac{n_1 n_2}{n(n-1)} \left(\sum_{i=1}^n c_i^2 \right) \quad (9)$$

as given, for example, in Prentice (1978).

Later researchers focused primarily on asymptotic normality as an expression of the full distribution theory represented by an i.i.d. sequence of the three vectors $\{L_i, R_i, T_i\}$ with $n_1/n \rightarrow p_1 \in (0, 1)$ as $n \rightarrow \infty$. Early work, such as Finkelstein (1986) and Fay (1996), which used likelihood methods but lacked rigorous proofs of asymptotic normality, considered the use of observed Fisher information. However, the presence of high dimension nuisance parameters in baseline survival and transformation g respectively, lead to information matrices of order $O(n)$ that complicated the computation of asymptotic variance.

Sun *et al.* (2005) provided rigorous proofs of asymptotic normality under assumptions that are more restrictive than required for validating permutation significances. The sequence $\{L_i, R_i, T_i\}$ is assumed to be i.i.d. with independent censoring, in which (L_i, R_i) is independent of T_i ; thus dependent censoring is not allowed and random censoring distributions are not dependent on i . Additionally, in their setup, random censoring leads to a trichotomy of data outcome types $(0, l_i]$, $(l_i, r_i]$, and (r_i, ∞) according to whether $T_i \leq L_i$, $T_i \in (L_i, R_i]$, or $T_i > R_i$. Thus all data are strictly interval-censored and exact survival observations cannot be considered. An asymptotic variance estimate, derived in Sun *et al.* (2005, thm. 1) and relevant for all statistics in class (4) with $\rho(0) = 0 = \rho(1)$, can be shown to reduce to

$$v_a = \frac{n_1 n_2}{n^2} \left(\sum_{i=1}^n c_i^2 \right) \quad c_i = \frac{\rho\{\hat{S}(l_i)\} - \rho\{\hat{S}(r_i)\}}{\hat{S}(l_i) - \hat{S}(r_i)}. \quad (10)$$

Here, the convention $\hat{S}(\infty) = 0$ applies to make $\{c_i\}$ well-defined. This simple reduction in form for the rather abstruse asymptotic variance given in Sun *et al.* (2005,

thm. 1) seems to have gone unrecognized in the literature.

The variances v_p and v_a are essentially the same even though the asymptotic theories upon which they are based are entirely different. The estimated Normal $(0, v_a/n)$ limit for U/\sqrt{n} in Sun *et al.* (2005) accounts for random variation in the triples $\{L_i, R_i, T_i\}$ but holding $\sum_i z_i = n_1$ in such a way that $n_1/n \rightarrow p_1 \in (0, 1)$. The estimated Normal $(0, v_a/n)$ limit makes use of the variations in $\{L_i, R_i, T_i\}$ but only through the permutational variation in U/\sqrt{n} . The essential reason for this equivalence is the assumption made in both approaches that censoring is independent of treatment allocation. The implications for this are: (i) The attained values $\{(l_i, r_i)\}$ are ancillary statistics upon which probability is conditioned; thus the weights $\{c_i\}$ are also ancillary and fixed when computing null variance. For inference purposes, the reference set for experimental repeatability under H_0 fixes the observed pattern of censoring. (ii) With fixed censoring pattern, only permutational variation in U/\sqrt{n} remains and can be judged as “asymptotically sufficient” for assessing the validity of H_0 .

Both asymptotic variances v_a and v_p (with n replacing $n - 1$) are simply the variance of $\sum_{i=1}^n c_i B_i$ where $\{B_i\}$ are assumed to be i.i.d. Bernoulli ($\hat{p}_1 = n_1/n$). If the permutation distribution for (Z_1, \dots, Z_n) is considered as a marginal distribution of the permutation distribution of (Z_1, \dots, Z_N) with $N \gg n$ and $\sum_{i=1}^N Z_i = Np_1$, then it should be clear that the marginal distribution of (Z_1, \dots, Z_n) converges weakly to an i.i.d. Bernoulli (p_1) vector as $N \rightarrow \infty$ which accounts for the dominant portion of asymptotic variation in U/\sqrt{n} .

In summary, the conditions under which permutation tests are justified are considerably broader than those under which asymptotic normality can be justified. However even when the latter conditions are justifiable, the same normal approximation would be used in determining significance. Furthermore, in small samples, the normal approximation itself is somewhat of a dubious assumption. Peto and Peto (1972) and Self and Grossman (1986) report inaccuracies due to lack of normality rather than inaccuracy of variance v_p .

4 Saddlepoint Approximation

When considering the permutation distribution of statistic U in (4), the sequence of z_i variables is generally assumed to have the distribution of a random permutation vector $\xi = (\xi_1, \dots, \xi_n)^T$ with n_1 ones and n_2 zeros. Thus any one of the distinct $\binom{n}{n_1}$ permutations for ξ is assumed to have probability $\binom{n}{n_1}^{-1}$. The null distribution for U is determined by its linear dependence on the permutation vector ξ . The fact that this dependence is linear in ξ leads to the following characterization that makes it amenable to double saddlepoint approximation as shown in Booth and Butler (1990).

Suppose that Z_1, \dots, Z_n are i.i.d. Bernoulli (θ) for any $\theta \in (0, 1)$. Then the conditional distribution of $Z = (Z_1, \dots, Z_n)^T$ given $\sum_{i=1}^n Z_i = n_1$ is the marginal permutation distribution for ξ . This provides the conditional characterization for the null distribution of U as

$$U = \sum_{i=1}^n c_i \xi_i \sim Y = \sum_{i=1}^n c_i Z_i \quad \text{given} \quad X = \sum_{i=1}^n Z_i = n_1.$$

If u_0 is the observed value of the statistic (2), the permutational mid- p -value

$$P(U < u_0) + \frac{1}{2}P(U = u_0) \simeq \hat{P}(Y \leq u_0 | X = n) \quad (11)$$

where the right side of (11) is the continuous version of the Skovgaard (1987) saddlepoint approximation for $P(Y \leq u_0 | X = n)$ given in Butler (2007, eqn. 4.12) and based on the joint MGF of (X, Y)

$$M_{X,Y}(s, t) = \prod_{i=1}^n \{1 - \theta + \theta \exp(s + c_i t)\};$$

see Butler (2007, ch. 6) for additional discussion on why the mid- p -value approximation in (11) holds.

5 Examples and simulations

Three real data examples of permutational rank tests are presented in the first three subsections. The fourth subsection presents a simulation study that compares the ac-

curacy of saddlepoint methods with asymptotic normal methods when approximating permutational mid- p -values.

The following abbreviations are used for the various test statistics given in (2) throughout: the Gehan-Mantel (GM) test using weights in (3), the log-rank (LR) and logistic-weighted (LW) tests in the class (4) with $\rho(y) = y \ln y$ and $\rho(y) = y^2 - y$ respectively, the Self-Grossman tests using the ‘‘Simple 2’’ option with extreme value (SG-E) and logistic (SG-L) errors respectively, the Finkelstein-Sun (FS) test in (8), and the test (SZZ) in Sun *et al.* (2005) using weight function $\rho(y) = (y \ln y)y(1 - y)$.

Executable programs for implementing all 7 tests using saddlepoint and normal approximations can be downloaded from the second author’s website.

5.1 Finkelstein and Wolfe (1985) data

Data consist of time until breast retraction for cancer patients who received lumpectomy. The control group of $n_2 = 46$ patients received radiotherapy while the treatment group of $n_1 = 48$ received radiotherapy supplemented with chemotherapy. One-sided mid- p -values and p -values are listed in Table 1 for the alternative hypothesis that breast retraction occurred sooner for the treatment group, e.g. $H_1 : S_1(t) < S_2(t)$.

Table 1. One-sided mid- p -value approximations for the Finkelstein and Wolfe (1985) data set with $n = 94$. Simulated p -values agreed with simulated mid- p -values to the accuracy displayed.

	Extreme-value weights			Logistic weights			
	LR	SG-E	FS	GM	LW	SG-L	SZZ
Using all 94 subjects:							
Sim. mid- p -value ¹	0.0033	0.0029	0.0034	0.0185	0.0149	0.0153	0.0001
Saddlept. Approx.	0.0034	0.0029	0.0036	0.0184	0.0148	0.0154	0.0001
Normal Approx.	0.0036	0.0030	0.0038	0.0187	0.0151	0.0157	0.0002

¹Based on 10^6 randomly generated permutations of treatment/control labels from the $\binom{n}{n_1}$ possible holding n_1 and n_2 fixed.

The saddlepoint approximation is somewhat closer to the simulated mid- p -values than the normal approximation that uses $N(0, v_p)$. Because of the relatively large sam-

ple size, the normal approximation is entirely adequate for the application. The three tests motivated by extreme-value weights show mid- p -values in the range 0.0029 – 0.0038 while those using logistic weights range from 0.0148 – 0.0187. By comparison, Fay (1996) reported two-sided normal mid- p -value approximations of 0.007 and 0.030 for LR and LW respectively, which agree with the values in Table 1. Finkelstein (1986) used an asymptotic variance based upon observed Fisher information and reported a two-sided normal approximation of 0.004 for either LR or FS (which one is unclear) that is somewhat smaller. Sun *et al.* (2005) report two-sided normal approximations for LR and SZZ as 0.007 and 0.0004 respective in agreement with Table 1.

5.2 Hoel and Walberg (1972) data

Data are current-status responses at death times where status is the (non)presence of a lung tumour for $n = 144$ RFM mice subjected to two treatments. A control group of $n_2 = 96$ mice was exposed to a conventional environment while a treatment group of $n_1 = 48$ mice experienced a germ-free environment. Table 2 provides mid- p -value approximations for the alternative $H_1 : S_1(t) > S_2(t)$. Except for SZZ, the

Table 2. One-sided mid- p -value approximations for Hoel and Walberg (1972) data. Simulated p -values agreed with simulated mid- p -values to the accuracy displayed.

	Extreme-value weights			Logistic weights			
	LR	SG-E	FS	GM	LW	SG-L	SZZ
Sim. mid- p -value ¹	0.1465	0.1256	0.1418	0.2048	0.1347	0.2827	0.1074
Saddlept. Approx.	0.1461	0.1252	0.1410	0.2044	0.1348	0.2821	0.1077
Normal Approx.	0.1455	0.1244	0.1405	0.2026	0.1341	0.2819	0.1075

¹Based on 10^6 randomly generated permutations of treatment/control labels from the $\binom{n}{n_1}$ possible holding n_1 and n_2 fixed.

saddlepoint approximation is more accurate although both approximations perform well for the large sample size. Finkelstein (1986) reported a two-sided normal approximation of 0.10 for LR using an asymptotic variance based upon observed Fisher information. This is much smaller than comparable value 2×0.1455 from Table 2.

5.3 Lindsey and Ryan (1998) data

AIDS data are given in Table II of Lindsey and Ryan (1998) who analyzed the original data given in Richman *et al.* (1990). Of interest is the time to development of drug resistance to zidovudine and its dependence on the stages of the disease, early or late, which define the treatment and control groups respectively. The data consist of $n_1 = 17$ ($n_2 = 14$) treatment (control) patients among which 7 (11) are interval-censored and 10 (3) are right-censored. Lindsey and Ryan (1998) comment that this is a challenging data set to analyze due to small sample sizes and the very wide intervals for interval censoring that resulted from infrequent periodic assessment due to associated costs.

Table 3 provides mid- p -values suggesting that late-stage patients show an earlier onset of drug resistance than early-stage patients. With smaller sample sizes, saddlepoint approximations show greater accuracy than normal approximations. Lindsey and Ryan (1998) found the computation of the FS statistic to be “unstable” and therefore did not report a p -value for this test in their Table IV summarizing the analysis.

Table 3. One-sided mid- p -value approximations for Hoel and Walberg (1972) data. Simulated p -values agreed with simulated mid- p -values to the accuracy displayed.

	Extreme-value weights			Logistic weights			
	LR	SG-E	FS	GM	LW	SG-L	SZZ
Sim. mid- p -value ¹	0.0016	0.0008	0.0016	0.0001	0.0009	0.0003	0.0014
Saddlept. Approx.	0.0018	0.0008	0.0016	0.0001	0.0010	0.0003	0.0012
Normal Approx.	0.0027	0.0011	0.0024	0.0004	0.0014	0.0005	0.0018

¹Based on 10^6 randomly generated permutations of treatment/control labels from the $\binom{n}{n_1}$ possible holding n_1 and n_2 fixed.

5.4 Simulations

The saddlepoint accuracy seen in Tables 1-3 occurs consistently over a wide range of conditions. For the LR and LW tests, a simulation study was conducted to determine

the accuracy of saddlepoint and normal mid- p -value approximations over balanced sample sizes of $n_1 = n_2 = 20, 35,$ and 50 ; using various proportions of partially interval-censored data; and using both logistic and extreme-value distributions for log-survival times. Simulation results from the respective error distributions are shown in Tables 4 and 5.

Each row in Table 4 summarizes the accuracy of saddlepoint and normal approximations over 1000 data sets generated in the following way. First the frequencies for the various types of partially interval-censored data were determined by using the percentage composition (% Comp.) in column 1 as multinomial percentages. For example, in the first row, the distribution frequencies of interval-censored; exact; left-censored; and right-censored observations in each treatment group were given as Multinomial $(20; 0.90, 0.0, 0.05, 0.05)$ as indicated by a % Comp. of 90;0;5;5. For control group, an exact survival time T_i was generated so $\ln T_i$ has a standard logistic distribution. An interval-censored value $(L_i, R_i]$ was generated as the interval values that bracket the exact log-logistic survival time T_i , which corresponds to regularly spaced follow up. A left-censored value $(0, L_i]$ was simulated so L_i was Uniform $(0, 5)$, and a right-censored value (R_i, ∞) had R_i as the greatest integer less than a simulated Uniform $(0, 5)$. The range $(0, 5)$ is 83.3% of the support for the log-logistic distribution. For the treatment group, survival times were shifted by an amount $\Delta > 0$ so that an exact time is $T_i + \Delta$, an interval-censored range consists of the integers bracketing $T_i + \Delta$, and left- and right-censored values $(0, L_i]$ and (R_i, ∞) had L_i and R_i based on a Uniform $(0, 5 + \Delta)$ simulation. The value of Δ was chosen so that the mean mid- p -value over the 1000 data sets was about 5% which is the most interesting situation.

Each row of Table 4 summaries 1000 of these data-set simulations. Column “Sad. Prop.” shows the percentage of the 1000 data sets for which the saddlepoint mid- p -value approximation was closer to the “true” mid- p -value than the normal approximation. For each data set, the “true” mid- p -value was determined through simulation of 10^6 random permutations of the treatment/control labels. The column “% Sad. Rel.

Err.” gives the average relative absolute error of the saddlepoint mid- p -value from the “true” mid- p -value expressed as a percent. For the log-rank test LR, saddlepoint errors ranged from 1.6 – 3.0% while for the logistic weighted test LW, relative error was 2.9 – 15.1%. As indicated in the table, saddlepoint approximations are mostly closer than normal approximations with substantially smaller percentage relative error. The normal approximations have relative errors that deteriorate with smaller sample sizes.

Table 5 shows the same sort of simulations but generating survival time T_i as an Exponential (1) distribution so $\ln T_i$ has an extreme value distribution. Left- and right-censored values were generated the same way. The range $(0, 5)$ is 99.3% of the support for the distribution of T_i . The same conclusions follow from Table 5.

Table 4. Simulations showing relative errors of mid- p -value approximations for the LR and LW tests using varying compositions of partially interval-censored data. Survival times were simulated as log-logistic.

% Comp.	LR			LW		
	Sad. Prop.	% Sad. Rel. Err.	% Nor. Rel. Err.	Sad. Prop.	% Sad. Rel. Err.	% Nor. Rel. Err.
$n_1 = n_2 = 20$						
90;0;5;5	95.4	2.97	169.8	96.9	9.39	77.9
70;3(10) ¹	93.9	2.15	119.9	96.2	6.73	56.8
40;3(20)	93.6	1.62	48.5	94.5	2.91	32.0
$n_1 = n_2 = 35$						
90;0;5;5	95.5	2.94	69.3	93.9	10.5	41.3
70;3(10)	94.5	2.75	48.5	92.9	9.79	35.6
40;3(20)	94.7	2.22	35.6	93.6	8.76	35.1
$n_1 = n_2 = 50$						
90;0;5;5	94.5	2.72	40.5	87.5	15.1	37.1
70;3(10)	95.2	2.42	34.4	89.3	12.4	33.2
40;3(20)	92.9	2.19	21.8	87.3	7.59	21.4

¹Denotes the multinomial percentages used for simulating the frequencies of interval-censored; exact; left-censored; and right censored observations. The notation 70;3(10) means 70;10;10;10.

Table 5. Relative errors as in Table 4 but with survival times simulated as Exponential (1).

% Comp.	LR			LW		
	Sad. Prop.	% Sad. Rel. Err.	% Nor. Rel. Err.	Sad. Prop.	% Sad. Rel. Err.	% Nor. Rel. Err.
$n_1 = n_2 = 20$						
90;0;5;5	85.5	7.03	247	82.5	10.6	51.6
70;3(10)	93.8	4.42	268	95.6	11.0	94.2
40;3(20)	94.8	1.51	58.3	95.0	3.81	40.0
$n_1 = n_2 = 35$						
90;0;5;5	93.2	2.92	51.0	86.0	5.59	20.4
70;3(10)	94.1	1.97	36.8	91.3	4.16	16.9
40;3(20)	94.3	1.96	27.2	93.8	6.56	31.4
$n_1 = n_2 = 50$						
90;0;5;5	93.8	2.18	30.2	78.6	4.60	12.6
70;3(10)	93.7	2.71	31.9	85.7	7.01	19.1
40;3(20)	90.7	2.24	17.7	87.6	7.50	21.2

6 Confidence interval for the treatment effect

Data from the pooled groups are assumed to be in the form $\{(\ln l_i, \ln r_i, z_i) : i = 1, \dots, n\}$ and on the log-scale. The significance of treatment effect δ is determined by subtracting δ from treatment intervals and testing the significance of the resulting data $\{(\ln \lambda_i - \delta z_i, \ln \rho_i - \delta z_i, z_i) : i = 1, \dots, n\}$. Suppose $\hat{S}(\delta)$ is the NPMLE of survival for the pooled data $\{(\ln \lambda_i - \delta z_i, \ln \rho_i - \delta z_i, z_i) : i = 1, \dots, n\}$ and $\hat{p}(\delta)$ is the one-sided saddlepoint mid- p -value for the LR or LW test of equal groups against alternative $H_1 : S_1(t) > S_2(t)$. Then, a $100(1 - \alpha)\%$ confidence interval for δ is $[\mathcal{L}, \mathcal{R}] = \{\delta : \alpha/2 \leq \hat{p}(\delta) \leq 1 - \alpha/2\}$. In practical applications, δ assumes values over a grid of increment 0.001 within range $[-B, B]$ for some $B > 0$.

A plot of $\hat{p}(\delta)$ vs. δ is a step function that can only change value when the increment $\delta \rightarrow \delta + 0.001$ results in a change in the structure of $(\ln l, \ln r)$ -bins within which $\hat{S}(\delta)$ places its mass. Such change can only occur when a treatment value, $\ln \lambda_{i_1} - \delta$ or $\ln \rho_{i_2} - \delta$, jumps over a control value, $\ln \rho_{i_3}$ or $\ln l_{i_4}$ respectively, with incremental change $\delta \rightarrow \delta + 0.001$. In applications, these plots have always been increasing however a proof for such is lacking. Table 6 shows 95% and 99% confidence

Table 6. Confidence intervals for the effect of adjuvant chemotherapy on the log-time scale for Finkelstein and Wolfe (1985) data.

Level	LR				LW			
	\mathcal{L}	\mathcal{R}	$\Delta\hat{p}(\mathcal{L})^1$	$\Delta\hat{p}(\mathcal{R})^2$	\mathcal{L}	\mathcal{R}	$\Delta\hat{p}(\mathcal{L})$	$\Delta\hat{p}(\mathcal{R})$
95%								
True	-0.8652	-0.1647	0.0017	0.0286	-0.8332	-0.0804	0.0185	0.0002
Sad.	-0.865	-0.165	0.0023	0.0285	-0.833	-0.076	0.0185	0.0006
Norm.	-0.865	-0.165	0.0022	0.0282	-0.833	-0.076	0.0183	0.0006
99%								
True	-1.0766	-0.0585	0.0082	0.0002	-0.8845	0.0515	0.0012	0.0019
Sad.	-1.129	-0.062	0.0014	0.0002	-0.885	0.052	0.0013	0.0019
Norm.	-1.129	-0.054	0.0015	0.0003	-0.885	0.053	0.0014	0.0012

¹Denotes the step height $\Delta\hat{p}(\mathcal{L}) = \hat{p}(\mathcal{L} + 0.001) - \hat{p}(\mathcal{L})$ at grid point \mathcal{L} computed according to the associated row; e.g. via simulation for the True row and via saddlepoint (normal) approximation for the Sad. (Norm.) row. ²Step height $\Delta\hat{p}(\mathcal{R}) = \hat{p}(\mathcal{R}) - \hat{p}(\mathcal{R} - 0.001)$.

intervals for δ computed by using $B = 3$. For all three intervals, endpoints \mathcal{L} and \mathcal{R} were determined so the interval $[\mathcal{L}, \mathcal{R}]$ is conservative with

$$\hat{p}(\mathcal{L}) \leq \alpha/2 < \hat{p}(\mathcal{L} + 0.001) \quad \text{and} \quad \hat{p}(\mathcal{R} - 0.001) < 1 - \alpha/2 \leq \hat{p}(\mathcal{R}). \quad (12)$$

Intervals in the Sad. row, obtained by inverting saddlepoint values $\hat{p}(\delta)$, are extremely accurate as compared to “True” intervals, obtained by determining each $p(\delta)$ using 10^6 random permutations of the treatment/control labels. Normal intervals (Norm.) agree with saddlepoint intervals at 95% level but differ at the 99% level. The reason for this agreement is due to large step sizes that occur in the tails of the step-function plots of $\hat{p}(\delta)$ for both saddlepoint and normal probabilities which share the same step locations; see the values $\Delta\hat{p}(\mathcal{L})$ and $\Delta\hat{p}(\mathcal{R})$ in Table 6. The pile up of mass occurring in the step-function plots of $\hat{p}(\delta)$ is a consequence of the pile up of mass in certain $(l, r]$ -bins that occur in the NPMLE $\hat{S}(\delta)$ as a result of interval censoring. Thus, while saddlepoint significance levels were seen to be considerably more accurate than their normal counterparts, this accuracy is not always converted into better or even different intervals with test inversion for the reasons described.

A $100(1 - \alpha)\%$ confidence interval for the percentage increase in mean (or median) treatment survival time over control survival time in the AFT model results by mapping $[\mathcal{L}, \mathcal{R}]$ through the function $\delta \rightarrow 100(e^\delta - 1)$. Inverting saddlepoint intervals for the LR test gives 95% and 99% confidence intervals $[-57.9, -15.2]$ and $[-67.7, -6.01]$ respectively. Inversion of the LW test gives respective intervals $[-56.5, -7.32]$ and $[-58.7, 5.34]$. A clinical interpretation of the first interval is that adjuvant chemotherapy reduces the mean (or median) time to breast retraction by 15.2% to 57.9% with confidence level 95%.

7 Algorithms for computation of NPMLE $\hat{S}(\delta)$

While permutation tests only require a single computation of \hat{S} , the NPMLE under H_0 , test inversion requires computation of $\hat{S}(\delta)$ for hundreds of potential choices of

treatment effect δ . Computation of $\hat{S}(\delta)$ must also be automated in an algorithm that is assured of converging to the requisite NPMLE. Both the EM algorithm, as outlined in Peto (1973) and Turnbull (1976), and the hybrid iterative convex minorant (hybrid ICM) in Wellner and Zhan (1997) failed in this automated computation of the NPMLE. The EM algorithm sometimes converged to local maxima, and was otherwise slow to converge. The hybrid ICM sometimes lead to cumulative sum diagrams that dropped below the x -axis thus resulting in negative probability estimates.

To deal with this, our EM-hybrid ICM algorithm runs 200 iterations of the EM algorithm starting with an initial estimate that places uniform mass in all $(l, r]$ -bins. If the resulting survival estimate is judged to be the NPMLE, the algorithm stops and uses the resulting survival estimate. If judged as not the NPMLE, the current survival estimate is used as the input for the hybrid ICM algorithm in Wellner and Zhan (1997). The hybrid ICM algorithm is stopped once the current estimate is judged to be the NPMLE. In all our programs a survival estimate is judged to be the NPMLE if it satisfies the Fenchel conditions given in Wellner and Zhan (1997 eqn 25) with error tolerance $\varepsilon = 10^{-7}$. Our EM-hybrid ICM algorithm that makes tandem usage of EM and hybrid ICM always converged to the NPMLE in our computations. This algorithm underlies all the executable programs provided when computing permutation significance levels and for inverting the LR and LW tests.

The essential problem encountered with the hybrid ICM algorithm occurred during the first few iterations. During these iterations, there is no guarantee that the first few cumulative sum diagrams of the form $\{(x_0, 0), (x_i, y_i) : i = 1, \dots, J\}$ will have all y -values that are nonnegative. If $\inf_{i \geq 1} y_i < 0$ then the ICM estimate for survival places negative probability into an $(l, r]$ -bin. This problem is removed by using our EM-hybrid ICM algorithm which uses an initial burn-in with the EM algorithm. To replicate this problem with the hybrid ICM, use the first 50 observations from the two treatment groups in DeGruttola and Lagakos (1989). Starting with an initial distribution that is uniform over all $(l, r]$ -bins, the second iteration leads to a

cumulative sum diagram with $y_1 < 0$.

References

- [1] Abd-Elfattah, E.F. & Butler, R.W. (2007). The weighted log-rank class of permutation tests: P -values and confidence intervals using saddlepoint approximations. *Biometrika* **94**, 543-551.
- [2] Booth, J.G. and Butler, R.W. (1990). Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika* **77**, 787-796.
- [3] Butler, R.W. (2007). *Saddlepoint Approximations with Applications*. Cambridge: Cambridge University Press.
- [4] Chung, K.L. (1974). *A Course in Probability Theory*. 2nd Ed. New York: Academic Press.
- [5] De Gruttola, V. & Lagakos, S.W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45**, 1-12.
- [6] Fay, M.P. (1996). Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics* **52**, 811-822.
- [7] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845-854.
- [8] Finkelstein, D.M. & Wolfe, R.A. (1986). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933-945.
- [9] Fleming, T. & Harrington, D.P. (1981). A class of hypothesis tests for one and two samples censored survival data. *Commun. Statist. A*, **10**, 763-794.
- [10] Gehan, E.A. (1965). A generalized two-sample Wilcoxon test for doubly-censored data. *Biometrika* **52**, 650-653.
- [11] Hoel, D.G. and Walburg, H.E. (1972). Statistical analysis of survival experiments. *J. Nat.Cancer Inst.*, **49**, 361-372.
- [12] Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd Ed. New York: Wiley.
- [13] Lindsey, J.C. & Ryan, L.M. (1998). Tutorial in biostatistics methods for interval-censored data. *Statist. Med.* **17**, 219-238.
- [14] Mantel, N. (1967). Ranking procedures for arbitrarily restricted observation. *Biometrics* **23**, 65-78.
- [15] Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics* **22**, 86-91.

- [16] Peto, R. & Peto, J. (1972). Asymptotic efficient rank invariant test procedures (with Discussion). *J. R. Statist. Soc. A*, **135**, 185-206.
- [17] Prentice, R.L. (1978). Linear rank tests with right-censored data. *Biometrika* **65**, 167-179.
- [18] Richman, D.D., Grimes, J.M. & Lagakos, S.W. (1990). Effect of stage of disease and drug dose on zidovudine susceptibilities of isolates of human immunodeficiency virus. *Journal of AIDS*. **3**, 743-746.
- [19] Skovgaard, I.M. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Prob.* **24**, 875-887.
- [20] Self, S.G. & Grossman, E.A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics* **42**, 521-530.
- [21] Sun, J. (1996). A non-parametric test for interval-censored data with application to AIDS studies. *Statist. Med.* **15**, 1387-1395.
- [22] Sun, J., Zhao, Q., & Zhao, X. (2005). Generalized log-rank tests for interval-censored failure time data. *Scand. J. Statist.* **32**, 49-57.
- [23] Turnbull, B.W. (1976). The empirical distribution with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc. B*, **38**, 290-295.
- [24] Wellner, J.A. & Zhan, Y. (1997). A hybrid algorithm for computation of the non-parametric maximum likelihood estimator from censored data. *J. Amer. Statist. Assoc.* **92**, 945-959.