The Canadian Journal of Statistics

Vol. xx, No. yy, 20??, Pages 1–25

La revue canadienne de statistique

Saddlepoint approximations for rank-invariant permutation tests and confidence intervals with interval-censoring

Ehab Abd-Elfattah[1] and Ronald Butler[2]*

[1]*Department of Mathematics, Faculty of Education, Ain Shams University, Cairo, Egypt*

[2]*Statistical Science, Southern Methodist University, Dallas, 75275-0332 USA*

Key words and phrases: Expectation maximization algorithm; interval censoring; iterative convex minorant algorithm; permutation test; saddlepoint approximation; two sample test.

MSC 2010: Primary 62N02; secondary 62N03

Abstract: Interval-censored data occur when subjects are assessed by using regular follow up. In such instances, we consider rank-invariant permutation tests to test the significance of a treatment versus a control. For a wide class of such tests, which includes the Peto & Peto class, we present saddlepoint approximations for the exact permutation mid-p-values which achieve extremely small relative errors. The speed and stability of these saddlepoint computations make them practicable for inverting the permutation tests and we compute nominal $100(1-\alpha)\%$ confidence intervals for the treatment effect. Such confidence intervals are of substantial clinical importance since, more than simply stating the level of statistical significance, they quantify the significant benefit of the treatment by providing a confidence interval for the percentage increase in mean (or median) treatment survival time as compares to control. Our methodology makes

heavy usage of nonparametric MLEs (NPMLEs) for survival functions and some limitations of existing algorithms, such as the hybrid ICM algorithm, are noted and accommodated. *The Canadian Journal of Statistics* xx: 1–25; 20?? © 20?? Statistical Society of Canada

Résumé: Insérer votre résumé ici. *La revue canadienne de statistique* xx: 1–25; 20?? © 20?? Société statistique du Canada

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, 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. We consider both data types as well as partly interval-censored data for which some exact survival times are observed. The data follow a two sample design commonly used in clinical trials in which a treatment group is compared with a control group.

To assess the significance of the treatment benefit, we consider a large class of rank-invariant permutation tests which includes seven tests already established in the literature and described in §2. For all the tests, we compute mid-p-values

* *Author to whom correspondence may be addressed.*

E-mail: rbutler@smu.edu

by using saddlepoint approximations for the associated null permutation distributions. The mid-p-values for these tests are simply their p-values subtracting half of the boundary probability for the observed cutoff value. The mid-p-value saddlepoint approximations provide extremely accurate analytical substitutes for exact permutation significance levels in both small and large samples and entail no simulation. For mid-p-values in the tails near the $2.5 - 5\%$ quantiles, approximation based on simulation can be time consuming since it requires reasonably large simulation sample sizes to replicate saddlepoint accuracy in terms of comparable relative error. Normal approximations offer quite adequate approximation to significance levels of exact permutation tests in large samples. However, in this and other applications, they are almost always less accurate than saddlepoint methods regardless of the sample size. Evidence for this is given in the simulations of §3.3.

These new methods extend the saddlepoint techniques, developed in Abd-Elfattah & 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.

The speed, accuracy, and stability of saddlepoint methods in determining permutation mid-p-values allow for the inversion of interval-censoring tests to determine arbitrary level confidence intervals for an assumed treatment effect δ. In an accelerated failure time (AFT) model, let δ be the treatment effect in log-time

and assume independent interval censoring. Also, if $\hat{p}(\delta)$ denotes the saddlepoint permutation significance for a one-tailed test of $H_0 : \delta = 0$, then we compute a $100(1-\alpha)\%$ confidence interval $[\mathcal{L}, \mathcal{R}]$ for δ which consists of those values of δ that are not significant at level $\alpha/2$ in each of the one-tailed tests, i.e.

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

The main benefit of such a confidence interval is that its image under the mapping $\delta \rightarrow 100(e^{\delta} - 1)$ provides a $100(1-\alpha)\%$ confidence interval for the percentage increase in median (or mean) treatment survival time over control survival time in the AFT model. Such percentage increases quantify the magnitude of the significant benefit and therefore convey the clinical importance of the treatment much more than just a statement of the significance level. Such intervals have not been reported in the literature presumably due to the complexity and intensity of the computations involved.

Implementation of these rank-invariant tests is complicated by the need to compute the nonparametric maximum likelihood estimate (NPMLE) for survival $\hat{S}(t)$ as discussed in Peto (1973) and Turnbull (1976). Indeed, confidence interval $[\mathcal{L}, \mathcal{R}]$ requires intensive use of such NPMLE computations since each $\hat{p}(\delta)$ is computed over a fine grid of thousands of δ-values and each $\hat{p}(\delta)$ requires a separate survival estimate denoted as $\hat{S}_{\delta}(t)$. Both the EM algorithm in Turnbull (1976) and the hybrid iterative convex minorant (hybrid ICM) algorithm in Wellner & Zhan (1997) failed to converge to the NPMLE at some point during the

course of our computations. Failure of the EM was expected since its iterates are only assured of converging to a local maximum. Failure, however, of the ICM algorithm was unexpected since its iterates have been proven to converge to a global maximum; see Wellner & Zhan (1997). 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 tandem in this way, we always achieved convergence to the NPMLE 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 seven tests considered. Additional programs also compute arbitrary $100(1-\alpha)\%$ confidence intervals $[\mathcal{L}, \mathcal{R}]$ for treatment effect based upon inverting the two most commonly used permutation tests.

To summarize, this paper makes two important contributions for inference in two sample designs subject to interval censoring. First, it provides saddlepoint approximations to compute permutation significance levels of treatment benefit. Unlike other methods, such approximations can be routinely used with the expectation that they are virtually exact in all situations involving both large or small sample sizes and with heavily or lightly censored data. Secondly, such tests are inverted to give $100(1-\alpha)\%$ confidence intervals for percentage increase in

median (mean) treatment survival time over control time. Such confidence intervals have not been considered in the literature. Our examples suggest that test inversion based on saddlepoint approximation leads to intervals that attain coverage levels in virtual agreement with their intended nominal levels.

The paper is organized as follows. Section 2 considers the class of rank invariant tests, discusses permutation significance versus asymptotic normal significance, and outlines how saddlepoint approximations are used to compute permutation significance. Section 3 considers three real data examples, simulations to assess the accuracy of saddlepoint and normal significance computations, and develops the test inversion for confidence interval determination. Section 4 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 $\{(l_i, r_i, z_i) : i = 1, \ldots, n\}$ where $(l_i, r_i]$ is the range of time within which the ith 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 we consider rank-based permutation tests for testing H_0 : $S_1(t) \equiv S_2(t) = S(t)$ versus the one-sided stochastically ordered alternative $H_1 : S_1(t) > S_2(t)$. 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 the tests reject H_0 for small U. If u is the observed value of U, then the attained one-sided mid-p-value is computed as $P(U < u) + P(U = u)/2$ under the assumption that U is uniformly distributed over all $\binom{n}{n_1}$ distinct permutations of its treatment labels $\{z_i\}$.

When testing H_0 versus the two-sided alternative, $H_1 : S_1(t) \neq S_2(t)$ for some t, the two-sided test rejects H_0 for sufficiently small or sufficiently large U. Such a test is justified by reversing the roles of the treatment and control groups and recognising that the resulting test rejects for small $\sum_{i=1}^{n} c_i(1 - z_i)$ or when U in (2) is large. Thus one option for a two-sided mid-p-value is to compute the smaller of the two values for $P(U < u) + P(U = u)/2$ and $P(U > u) + P(U = u)/2$ and double it. This corresponds to the two-sided p-value assigned by an asymptotic normal approximation to the null permutation distribution.

Peto & Peto (1972, §4) proposed a subclass of such generalized rank-invariant tests with the form

$$U = U(\hat{S}) = \sum_{i=1}^{n} z_i \frac{\rho\{\hat{S}(l_i)\} - \rho\{\hat{S}(r_i)\}}{\hat{S}(l_i) - \hat{S}(r_i)}, \tag{3}$$

where the weight function ρ, defined on $[0, 1]$, determines the specific test. Survival estimate $\hat{S}$ is the NPMLE of the survival function S under the null hypothesis computed by pooling the set of interval-censored data $\{(l_i, r_i] : i = 1, \ldots, n\}$ and making the assumption that $\hat{S}(\infty) = 0$ in (3). Also, for U to be meaningfully defined, only weight functions ρ are considered for which $\rho(0) = 0 = \rho(1)$ are defined by continuity. We consider three members of the Peto & Peto subclass (3) as listed in Table 1.

Table 1: Three rank-invariant tests to be considered from the Peto & Peto subclass along with their weight function ρ.

Name	Symbol	$\rho(y)$
log-rank	LR	$y \ln y$
logistic-weighted	LW	$y^2 - y$
Sun et al. (2005)	SZZ	$(y \ln y)y(1 - y)$

Four additional tests from the more general class (2) but not in the Peto & Peto class are also considered. They include (i) GM, a Wilcoxon-type test proposed by Gehan (1965) and Mantel (1967); (ii) and (iii) SG-E and SG-L, tests from Self & Grossman (1986) that use their "Simple 2" option and are motivated from AFT models with extreme minimum value errors and logistic errors respectively; and (iv) FS, a test proposed by Finkelstein (1986, eqn. 12) and Sun (1996) whose rank score weights are based on imputations of the EM algorithm

when used to compute NPMLE $\hat{S}$. Zhao & Sun (2004) have generalized this test to accommodate partly interval-censored data. Explicit expressions for the rank score weights of all four tests are given in the Supplementary Materials.

Based upon the underlying motivation in the development of the seven tests, we can expect tests LR, SG-E, and FS to provide exceptional power against locational shifts in an accelerated failure time (AFT) model with extreme minimum value errors. Furthermore, tests LW, GM, and SG-L should demonstrate good power for detecting locational shifts in an AFT model with logistic errors. The weights for SZZ have been chosen to be the odd test out in the group.

2.1. Permutation and 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 & Peto (1972), as well as later researchers such as Self & Grossman (1986) and Fay (1996).

Both permutation and asymptotic normal significances require independent censoring mechanisms for their validity as well as censoring mechanisms that do not depend upon group membership. Beyond this, however, the requirements for permutation significances are quite weak and only require "balanced" censoring in the two groups which is generally achieved with randomized assignment of subjects to groups. Thus, if $\{L_i, R_i, Z_i\}$ represent the observables for subjects, then randomized assignment ensures that $\{L_i, R_i\}$ is independent of $\{Z_i\}$. When this is not the case, then permutation methods may not be valid as discussed in

Fay and Shih (2012, §3) who provide examples in which assessment times L_i, R_i are allowed to depend on group membership Z_i. However, assuming randomized assignment of subjects to groups, then permutation significances still allow for joint distributional dependence of L_i, R_i on the index i or covariate(s) y_i associated with subject i. Such dependence allows for the possibility that individual subjects are censored according to their individual attributes and was discussed extensively by Mantel (1967) who noted that heterogeneity in the distributions of $\{L_i, R_i\}$ with i or y_i balances out in the two groups with sufficiently large samples. Such heterogeneity can, however, reduce the power of the test.

Asymptotic normal significances rely on proofs that make more restrictive assumptions than are required for validating permutation significances. For example, Sun et al. (2005) provided rigourous proofs that a standardized U is asymptotically $N(0, v^2)$ with v^2 given in their Theorem 1. This applies for test statistics in the Peto & Peto subclass (3) under case II independent censoring (Sun, 2006, p. 11-15) when the sequence $\{L_i, R_i, T_i\}$ is i.i.d. under H_0, i.e. when treatment and control groups have a common survival distribution. Thus, formally, censoring distributions cannot depend on i or on an associated covariate value y_i, however one suspects that such restrictive conditions can be relaxed to allow more diverse censoring conditions. To accommodate exact survival observations, Zhao et al. (2008) extended these asymptotic normal results thus allowing normal approximation theory to apply to partly interval censored data. Oller & Gómez (2012) showed that such normal limits agree with the standard normal approximations for the permutation distributions as given in Prentice (1978) so that

$$v^2 = v_p^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} c_i^2\right)\sum_{j=1}^{n}(z_j - \bar{z})^2 = \frac{n_1 n_2}{n(n-1)}\sum_{i=1}^{n} c_i^2 \qquad (4)$$

where v_p has been given in Prentice (1978).

Perhaps the main point to be made here is that saddlepoint approximations will be seen as providing virtually exact computation of permutation significances in all settings for which permutation tests are valid including settings in which censoring may depend upon the individual. For example, assessment times L_i, R_i could depend on a covariate y_i such as gender or some other subpopulation designation. Of course the independent censoring (Sun, 2006, p. 11-15) assumption must now apply conditional upon such covariates.

2.2. Software packages for p-value computation

Four existing software packages can be used to compute p-values for treatment significance with (partly) interval-censored data. The four packages are listed in the rows of Table 2 and include three R packages and SAS (So, Johnson, & Kim, 2010). The R software includes the "interval" package (Fay & Shaw, 2010), the "glrt" package (Zhao & Sun, 2010), and the "FHtest" package (Oller & Langohr, 2013). The various capabilities of the software are listed as the columns of Table 2. Such capabilities include the analysis of partly censored data (Part cens); p-value computation for 1- and 2-sided tests (Tails 1 and 2); p-value distributions based upon asymptotic normality (Norm), Monte Carlo (MC), and an exact network algorithm (Exact); and test statistics which can be computed (LR, FS, LW, FH_1 and FH_2). The tests indicated by FH_1 and FH_2 represent two different classes of generalized Fleming & Harrington (1981) tests. The SSZ statistics is a member of the FH_1 class considered by Sun et al. (2005) and Zhao et al. (2008) with weight function $\rho(y) = (y \ln y) y^\eta (1-y)^\lambda$ for $\eta \geq 0 \leq \lambda$. The other class FH_2 is due to Oller & Gómez (2012) and removes the factor $\ln y$ with weight function $\rho(y) = y^\eta (1-y)^\lambda$ for $\eta \geq 0 \leq \lambda$.

Our software computes mid-p-values using saddlepoint approximations which have distinct advantages over these existing computational packages. In particular, the Normal, MC, and Exact methods for computation are deficient respectively in accuracy, speed of computation, and range of applicability as compares to saddlepoint methods.

Table 2. Listed capabilities for the four computational packages are indicated by y (yes) with blank entries indicating no.

	Part	Tails		Computations			Test statistic coverage				
Package	cens	1	2	Norm	MC	Exact	LR	FS	LW	FH_1^*	FH_2^*
interval	y	y	y	y	y	y	y	y	y		
glrt	y		y	y			y	y		y	
FHtest	y	y	y	y	y	y	y		y		y
SAS	y		y	y			y	y			

Part cens, handles partly censored data; Tails, indicates 1- or 2-sided testing, Norm, asymptotic normal p-values; MC, Monte Carlo p-values; Exact, exact network algorithm p-values. $^*FH_1$ and FH_2 represent two different classes of generalized Fleming & Harrington (1981) tests as described in the text.

Furthermore, none of the four packages support the computation of GM, SG-E, or SG-L with Fay and Shaw (2010, §2.4) suggesting that the Self & Grossman (1986) procedures are too difficult to calculate. In addition, none of these packages undertake the challenging computations for inverting these tests to determine confidence intervals for the amount of treatment benefit.

2.3. Saddlepoint approximation

When considering the permutation distribution of statistic U in (2) or (3), the sequence of z_i variables is assumed to have the distribution of a random permutation vector $\xi = (\xi_1, ..., \xi_n)^T$ consisting of 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 $U = \sum_{i=1}^{n} c_i \xi_i$ 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 saddlepoint approximation as shown in Skovgaard (1987).

Suppose that $Z_1, ..., Z_n$ are i.i.d. Bernoulli (θ) for any $\theta \in (0, 1)$. Then the conditional distribution of $Z = (Z_1, ..., Z_n)^T$ given $\sum_{i=1}^{n} Z_i = n_1$ is the marginal permutation distribution for ξ. Thus, for fixed n_1, the marginal null distribution of $U = \sum_{i=1}^{n} c_i \xi_i$ is the conditional distribution of $Y = \sum_{i=1}^{n} c_i Z_i$ given $X = \sum_{i=1}^{n} Z_i = n_1$. From this equivalence, if u_0 is the observed value of the statistic (2) or (3), then the permutation mid-p-value is

$$P(U < u_0) + \tfrac{1}{2} P(U = u_0) = P(Y < u_0 | \, X = n_1) + \tfrac{1}{2} P(Y = u_0 | \, X = n_1). \tag{5}$$

This distributional equivalence is needed because the right side of (5) is amenable to saddlepoint approximation whereas the left side is not. To understand the nature of this approximation, we need to recognize two important facts: First, for given values of u_0 and n_1, the conditional probability on the right side of (5) is uniquely determined from the joint moment generating function (MGF) of (X, Y). Secondly, this MGF is simple to compute from the i.i.d. Bernoulli (θ) as

$$M_{X,Y}(s, t) = \prod_{i=1}^{n} \{1 - \theta + \theta \exp(s + c_i t)\}. \tag{6}$$

The Skovgaard (1987) saddlepoint approximation now provides the means by which $M_{X,Y}(s, t)$ and values u_0 and n_1 can be converted into an approximation for the right side of (5). Details of this are straightforward saddlepoint theory and are given in the Supplementary Material online.

3. EXAMPLES, SIMULATIONS, AND CONFIDENCE INTERVALS

Three standard data examples have been used almost exclusively to illustrate two sample tests with interval or current status data. In §3.1, we use these three data sets to show the performance of saddlepoint and normal approximations for the various permutation rank tests. In §3.2 we describe a simulation study that compares the accuracy of saddlepoint methods with asymptotic normal methods and present the numerical results in §3.3. The saddlepoint accuracy achieved in the simulations is discussed in §3.4. Finally, in §3.5, we invert the LR and LW tests using both saddlepoint methods and asymptotic normal methods to determine confidence intervals for the treatment effect.

To judge the accuracy of saddlepoint and normal approximations, an "exact" permutation significance is needed to make comparisons. Since such computation is well beyond the capabilities of the network algorithm in the "interval" package of R with larger sample sizes, we have chosen to use Monte Carlo mid-p-values based upon 10^6 permutations as surrogates for exact computations in our comparisons throughout. In order to check the accuracy of such a strategy, we used the network algorithm in R to compute an exact one-sided p-value for the LW test statistic applied to the first ten observations $(n_1 = 10 = n_2)$ of the breast cosmesis data of Finkelstein & Wolfe (1986) as described below. The resulting exact p-value of 0.48655 can be compared to a Monte Carlo p-value of 0.4864 ± 0.00098 based upon 10^6 permutations where the error provides a 95% confidence interval. The attained relative error is -0.031% and indicates an acceptable level of error to use in the accuracy assessments.

3.1. Example data sets

The breast retraction data of Finkelstein & Wolfe (1986) consist entirely of interval censored data. Control group 2 received radiotherapy while treatment group 1 received radiotherapy supplemented with chemotherapy. The alternative hypothesis is $H_1 : S_1(t) < S_2(t)$, that retraction time was shortened with adjuvant chemotherapy. One-sided mid-p values for the seven tests using saddlepoint and normal approximations are displayed in the upper portion of Table 3 along with Monte Carlo mid-p-values.

The saddlepoint approximation is somewhat closer to the simulated mid-p-values than the normal approximation that uses $N(0, v_p^2)$ with v_p^2 given in (4). Because of the relatively large sample sizes, 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 $0.0148 - 0.0187$. The values are consistent with p-values reported by Fay (1996) and Sun et al. (2005).

The saddlepoint and normal approximations for the seven tests are standard output of our downloadable executable program which executed in about 0.67 seconds for each pair of rows. Most of this time was used to compute the NPMLE $\hat{S}$ of the survival function for the pooled data, a fact also noted by Fay and Shaw (2010, §4.2). By comparison, the Monte Carlo mid-p-values in each row required about 5.04 seconds of execution time using a separate executable program.

The lung tumor data from Hoel & Walberg (1972) are current-status responses at death times where status is the (non)presence of a lung tumor for $n = 144$ RFM mice subjected to two treatments. The middle portion of Table 3 provides mid-p-value approximations for the alternative $H_1 : S_1(t) > S_2(t)$. Except for SZZ, the saddlepoint approximation is more accurate although both approxima-

tions perform well for the large sample sizes.

Table 3: One-sided mid-p-value approximations for the three data sets listed. 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
Breast cosmesis data				$n_1 = 48$		$n_2 = 46$	
Sim. mid-p-value[a]	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
Lung tumor data				$n_1 = 48$		$n_2 = 96$	
Sim. mid-p-value[a]	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
AIDs data				$n_1 = 17$		$n_2 = 14$	
Sim. mid-p-value[a]	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

[a]Based 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 AIDs data were taken from Table II of Lindsey & Ryan (1998) who analyzed the original data given in Richman, Grimes, & Lagakos (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 treatment (control) patients among which 7 (11) are interval-censored and 10 (3) are right-censored. Lindsey & Ryan (1998) note 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. The bottom of Table 3 provides mid-p-values

for $H_1 : S_1(t) > S_2(t)$ suggesting that late-stage patients show an earlier onset of drug resistance than early-stage patients. With such smaller sample sizes, saddlepoint approximations show greater accuracy than normal approximations.

3.2. Simulation Study

The saddlepoint accuracy seen in Table 3 occurs consistently over a wide range of conditions. For the log-rank (LR) and logistic-weighted (LW) tests, we conducted a simulation study to determine the accuracy of saddlepoint and normal mid-p-value approximations over balanced sample sizes of $n_1 = n_2 = 20,\ 40,$ and 80 and over unbalanced samples with $n_1 = 80$ and $n_2 = 40$; using various proportions of partly 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 the tables represents a particular setting for n_1 and n_2 along with a particular choice for the proportions of the various types of partly interval-censored data. For example in row 2 of Table 4, 1000 data sets were simulated with $n_1 = n_2 = 20$. Each observation in each data set could be interval-censored (due to periodic follow up) with probability (w.p.) 0.7, exactly observed w.p. 0.1, or left- or right-censored as a current status response (due to a single assessment) w.p. 0.2. Control survival times were simulated so $\ln T$ is standard logistic (with mean zero and variance 1) while treatment survival times were shifted upward on the time scale by the amount $\beta = 1.7$ as shown in the second column. The value of β was chosen so that the mean mid-p-value over the 1000 data sets was in the range $2.5 - 5\%$. The primary reason for using such β values is to show the saddlepoint accuracy at the boundary of significance which is the setting in which mid-p-value accuracy is most important. A second reason is that such

choice will allow us to assess the coverage accuracy for treatment effect when such tests are inverted as described in §3.5. The specific algorithm for simulating partly censored data, as it concerns row 2 of Table 3, is as follows:

1. Simulate $\mathbf{M} = (M_1, M_2, M_3) \sim$ Multinomial $(1; p_1, \ldots, p_3)$ such that $p_1 + p_2 + p_3 = 1$. In our example $p_1 = 0.7$, $p_2 = 0.1$, and $p_3 = 0.2$.

2. For a control group simulation, simulate $\ln T$ as standard logistic (with mean zero and variance 1) for Table 4. For Table 5, simulate T as Exponential (1). (a) If $M_1 = 1$, then T is interval censored by taking $L = \lfloor T \rfloor$ and $R = \lfloor T \rfloor + 1$ where $\lfloor T \rfloor$ denotes the greatest integer less than or equal to T. (b) If $M_2 = 1$, then take $L = T^-$ and $R = T$. (c) If $M_3 = 1$, then simulate random assessment time V so that $\ln V$ is Normal $(0, 1)$. If $V \leq T$, then take $L = V$ and $R = \infty$ so we have right-censoring, if $V > T$ then take $L = 0$ and $R = V$ so we have left-censoring.

3. For a treatment group simulation, replace T with $T + \beta$ and proceed as in 2(a)-(c) but replacing T with $T + \beta$.

In this algorithm, vector $\mathbf{M}$ designates the type of censoring, i.e. whether it is interval-censored based on periodic follow up $(M_1 = 1)$, not censored $(M_2 = 1)$, or a current status response $(M_3 = 1)$ leading to left- or right-censoring. Left- and right-censoring occur with equal frequency in Table 4 since both $\ln V$ and $\ln T$ are symmetric so $\ln V - \ln T$ is symmetry about zero to make $P(V > T) = 0.5$. This is not the case in Table 5 where $\ln T$ has an extreme value distribution which is skewed to the left and which makes left-censoring more likely. For a control simulation, $P(V > T) = 0.618$ and 61.8% tend to be left-censored; for a treatment simulation, this percentage is lower and dependent on the β-value used for the treatment shift.

Each subject in the simulation has an associated random vector $\mathbf{M}$ whose value indicates the way in which the subject is censored but also could be interpreted as the subpopulation from which the subject is taken. As such, the vector could be considered a covariate assigned to the subject thus making the censoring subject–dependent. As previously mentioned in §2.1, permutation tests remain valid with such dependence.

3.3. Simulation Results

The table entries summarize the accuracy of saddlepoint and normal approximations when compared with "exact" mid-p-values. In each row, 1000 data sets were generated as described in §3.2 and for each data set the two mid-p-value approximations were compared to an "exact" or Monte Carlo mid-p-value determined from simulation of 10^6 random permutations of the treatment/control labels. (In the Supplementary Material online, we justify the use of such a Monte Carlo value in place of the exact mid-p-value where it is shown to have a standard error of 0.00022. Additionally, we show that the tabulated relative errors are not affected by such substitution.) The column "Sad. Prop." shows the percentage of the 1000 data sets for which the saddlepoint mid-p-value approximation was closer to the Monte Carlo mid-p-value than the normal approximation. In Table 4, for both the LR test and the LW test, the saddlepoint approximation was most often closer particularly with smaller sample sizes. The column "% Sad. Rel. Err." ("% Nor. Rel. Err.") gives the average absolute relative error of the saddlepoint (normal) mid-p-value from the Monte Carlo mid-p-value when expressed as a percent. With log-logistic errors, relative errors of saddlepoint approximations range from $0.9 - 2.7\%$ for the LR test and $4.2 - 14.1\%$ for the LW test; comparable errors for the normal approximations are $16.0 - 123.5\%$ and $11.9 - 50.6\%$ respectively. The saddlepoint approximations show substantially

smaller percentage relative error overall than the normal approximations which, unlike the saddlepoint approximations, have relative errors that deteriorate with smaller sample sizes as well as with further deviation into the distributional tail.

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

		LR			LW		
% Comp.	β	Sad. Prop.	% Sad. Rel. Err.	% Nor. Rel. Err.	Sad. Prop.	% Sad. Rel. Err.	% Nor. Rel. Err.
			$n_1 = n_2 = 20$				
100;0;0[a]	1.7[b]	95.2	2.7	123.5	90.0	6.6	50.6
70;10;20	1.7	93.8	2.3	104.4	95.9	4.5	42.0
40;20;40	1.7	94.3	2.5	99.5	96.2	4.2	43.2
			$n_1 = n_2 = 40$				
100;0;0	1.5	93.0	2.6	62.0	90.8	14.1	42.8
70;10;20	1.3	94.9	2.7	49.0	92.4	10.0	35.6
40;20;40	1.1	92.9	2.5	44.8	93.1	8.2	33.3
			$n_1 = 80, n_2 = 40$				
100;0;0	0.9	97.3	1.4	73.7	91.1	8.2	27.4
70;10;20	0.9	96.6	2.0	74.9	88.3	6.6	21.1
40;20;40	0.8	96.9	1.5	65.7	82.3	5.6	15.3
			$n_1 = n_2 = 80$				
100;0;0	0.8	92.4	0.9	9.9	73.2	7.2	11.9
70;10;20	0.8	93.0	1.6	16.0	76.2	8.6	15.3
40;20;40	0.75	93.0	1.9	16.9	79.1	7.9	14.8

[a]Denotes the multinomial percentages used for simulating the frequencies of interval-censored, exact, and current status responses. [b]Treatment effect on the time scale in the simulation.

Table 5 shows the same sort of simulations but with control group survival times T_i generated from an Exponential (1) distribution so $\ln T_i$ has an extreme value distribution. Treatment group survival times were again shifted upward by β on the time scale. Saddlepoint approximations were most often closer to the Monte Carlo mid-p-values with relative errors in the range $1.6 - 15.0\%$ and

$3.2 - 18.1\%$ for the LR and LW tests respectively; the normal approximations by comparison have relative errors $16.9 - 390.3\%$ and $6.1 - 70.1\%$. One might expect the normal approximations to perform poorly here with smaller sample sizes since extreme value errors are left-skewed and not symmetric as in the log-logistic setting.

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

		LR			LW		
% Comp.	β	Sad. Prop.	% Sad. Rel. Err.	% Nor. Rel. Err.	Sad. Prop.	% Sad. Rel. Err.	% Nor. Rel. Err.
			$n_1 = n_2 = 20$				
100;0;0	0.85	81.1	15.0	390.3	77.3	18.1	70.1
70;10;20	0.9	96.3	4.0	266.8	94.3	7.3	63.2
40;20;40	0.8	97.2	2.5	142.3	94.6	3.2	33.6
			$n_1 = n_2 = 40$				
100;0;0	0.7	86.0	6.0	83.8	78.9	9.7	26.5
70;10;20	0.7	95.0	5.5	109.1	91.2	7.2	25.3
40;20;40	0.55	94.6	1.6	38.5	88.9	3.4	13.7
			$n_1 = 80, n_2 = 40$				
100;0;0	0.6	94.0	5.8	205.1	86.2	8.4	29.4
70;10;20	0.5	98.1	1.9	107.3	91.4	3.2	15.4
40;20;40	0.55	97.8	2.7	135.2	91.6	6.1	24.0
			$n_1 = n_2 = 80$				
100;0;0	0.5	88.1	3.1	23.4	72.3	6.9	11.5
70;10;20	0.4	92.2	1.6	16.9	71.3	3.3	6.1
40;20;40	0.45	91.5	2.7	26.0	79.0	5.9	12.3

3.4. Discussion of saddlepoint accuracy in the simulations

The simulations demonstrate the accuracy that saddlepoint approximations can achieve when approximating an empirical distribution such as a permutation or bootstrap distribution. Such accuracy has already been well established in a substantial body of literature which includes Robinson (1982), Davison & Hinkley

(1988, 1997), Feuerverger (1989), and Butler & Bronson (2002, 2012). Basically, if the permutation distribution of U is "coarsely" distributed, with large masses on a few points, then the accuracy from fitting a smooth saddlepoint approximation is likely to suffer; alternatively it will likely improve when the permutation distribution is "finer" and has small masses distributed over more points.

Our use of periodic follow up for censoring in the simulation presents perhaps the most challenging setting for such accuracy since it leads to a coarse permutation distribution for U. To see this, consider one of the simulations in which there is 100% interval censoring so that all interval-censored observations take the form $(j-1, j]$ for integer j with many "tied" values and j ranges over $j = 1, \ldots, N$ for the pooled data. The NPMLE $\hat{S}(t)$ is characterized by the finite values $\{\hat{S}(j) : j = 1, \ldots, N-1\}$. If U^* denotes a permuted value of U in (3) with randomized treatment labels $\{z_i\}$ which result in m_j treatment values that are assigned to range $(j-1, j]$, then

$$U^* = \sum_{j=1}^{N} m_j \frac{\rho\{\hat{S}(j)\} - \rho\{\hat{S}(j-1)\}}{\hat{S}(j) - \hat{S}(j-1)}.$$

Note that all of the $\binom{n_1}{m_1,\ldots,m_N}$ possible permutations out of the $\binom{n_1+n_2}{n_1}$ total that allocate m_j treatment labels to $(j-1, j]$ for $j = 1, \ldots, N$ lead to the same permuted value of U^* and this contributes to the coarseness of the permutation distribution with 100% interval censoring. At the other extreme, when no data are censored or tied, then such tied values of U^* are unlikely and there can be as many as $\binom{n_1+n_2}{n_1}$ distinct permuted values of U^* which makes the permutation distribution considerably finer.

This discussion explains some of the saddlepoint accuracy seen in Tables 4 and 5. Accuracy tends to be the worst with 100% interval–censoring (100;0;0 % Comp.) and best with the most uncensored observations (40;20;40 % Comp.).

A referee has suggested that other simulation schemes should be considered that incorporate random overlapping of intervals and random interval lengths. From the discussion above, it should be clear that such censoring mechanisms will result in a "finer" distribution of mass for NPMLE $\hat{S}(t)$ which in turn will lead to a finer permutation distribution for U. Thus, in such contexts, even greater saddlepoint accuracy can be expected than is given in Tables 4 and 5.

3.5. Confidence interval for the treatment effect

Let the data from the pooled groups assume the form $\{(\ln l_i, \ln r_i, z_i) : i = 1, \ldots, n\}$ on the log-scale. A treatment effect δ (on the log-scale) is a meaningful parameter in an AFT model that assumes $\ln T_i = \delta z_i + \varepsilon_i$ with $\{\varepsilon_i\}$ as i.i.d. error responses from a continuous distribution. We also assume independent interval censoring (Sun, 2006, §1.3.5) of any type so that censoring bounds provide no further information about survival times other than the bounds on their values. Under such assumptions, the δ-translated treatment intervals are indistinguishable from untranslated control intervals in the sense that they represent censoring from a single common control distribution. Thus, the intervals in $\{(\ln l_i - \delta z_i, \ln r_i - \delta z_i) : i = 1, \ldots, n\}$ represent interval censored data coming from a common control distribution. Detailed arguments for this assertion are given in the Supplementary Material online.

Such indistinguishability ensures that the data set $\{(\ln l_i - \delta z_i, \ln r_i - \delta z_i, z_i) : i = 1, \ldots, n\}$ satisfies the null hypothesis when δ is the true treatment effect in the AFT model. The significance of the value δ can be assessed by using a permutation test that this data follow from a common control distribution. In performing this test, suppose $\hat{S}_\delta$ is the NPMLE of survival for the pooled data $\{(\ln l_i - \delta z_i, \ln r_i - \delta z_i) : i = 1, \ldots, n\}$ and $\hat{p}(\delta)$ is the one-sided saddle-

point mid-p-value for the LR or LW test of H_0 versus $H_1 : S_1(t) > S_2(t)$. 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 $[-B_1, B_2]$ for some $B_1 > 0 < B_2$.

A plot of $\hat{p}(\delta)$ vs. δ is a step function that can only change value when the increment $\delta \longrightarrow \delta + 0.001$ 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, $\ln l_{i_1} - \delta$ or $\ln r_{i_2} - \delta$, jumps over a control value, $\ln r_{i_3}$ or $\ln l_{i_4}$ respectively, with incremental change $\delta \longrightarrow \delta + 0.001$. In applications, these plots have always been increasing however a proof for such is lacking. See Figure 1 in Abd-Elfattah & Butler (2007) for a similar plot with right-censoring.

Table 6: Confidence intervals for the effect of adjuvant chemotherapy on the log-time scale for breast cosmesis data.

	LR				LW			
Method	$\mathcal{L}$	$\mathcal{R}$	$\Delta\hat{p}(\mathcal{L})^a$	$\Delta\hat{p}(\mathcal{R})^b$	$\mathcal{L}$	$\mathcal{R}$	$\Delta\hat{p}(\mathcal{L})$	$\Delta\hat{p}(\mathcal{R})$
				95% level				
Exact	-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% level				
Exact	-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

[a]Denotes 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 Exact row and via saddlepoint (normal) approximation for the Sad. (Norm.) row.

[b]Step height $\Delta\hat{p}(\mathcal{R}) = \hat{p}(\mathcal{R}) - \hat{p}(\mathcal{R} - 0.001)$.

Table 6 shows 95% and 99% confidence intervals for δ computed over $[-1.5, 1]$ with incremental change 0.001 using the breast cosmesis data. For all three intervals, endpoints $\mathcal{L}$ and $\mathcal{R}$ were determined so the interval $[\mathcal{L}, \mathcal{R}]$ is

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

Intervals in the Sad. row, obtained by inverting saddlepoint values $\hat{p}(\delta)$, are extremely accurate when compared to "Exact" 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 the 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.

The inversion of such permutation tests has substantial clinical importance because it allows for more than simply a statement that "the treatment group had a significantly shorter recovery time than the control group." It rather allows "shorter" to be quantified by providing a $100(1-\alpha)\%$ confidence interval for the percentage decrease in mean (or median) recovery time for treatment versus control group in the AFT model. These confidence intervals are computed by mapping $[\mathcal{L}, \mathcal{R}]$ through the function $\delta \to 100(e^\delta - 1)$. Thus, for the LR test, this gives $[-57.9, -15.2]$ and $[-67.7, -6.01]$ as 95% and 99% confidence intervals respectively. From the first interval we may conclude that adjuvant chemotherapy reduces the mean (or median) time to breast retraction by 15.2% to 57.9% with confidence level 95%. Inversion of the LW test gives $[-56.5, -7.32]$ and

$[-58.7, 5.34]$ as the respective 95% and 99% confidence intervals.

Our downloadable executable file which inverts saddlepoint and normal tests for the LR and LW statistics required 14.0 minutes to determine the 99% confidence intervals in Table 6 using a grid for δ of increment 0.001 over the range $[-1.5, 1]$. By comparison, 105 minutes were required to determine the "exact" intervals using a different executable simulation program.

The coverage accuracy attained by the saddlepoint and normal intervals can be assessed using the relative error entries computed in Tables 4 and 5. The analysis in the Appendix shows that such an assessment is possible if we assume that saddlepoint relative error is roughly the same in the left and right tails of the null permutation distribution for log-survival times. In particular, this error analysis shows that a saddlepoint relative error of $R\% = 2.5\%$, as in the Sad. LR entry for Table 4 row 6, leads to an absolute error of coverage of $R\%/20 = 2.5\%/20 = 0.12\%$ as indicated in Table 7. When using saddlepoint tests, the largest error in coverage is therefore 0.41% for the LW test and quite small while for normal tests the largest coverage error is 2.24% for the LR test.

Table 7. Percentage coverage error when inverting saddlepoint (Sad.) and normal (Nor.) permutation tests for 95% confidence intervals of the treatment effect. The two settings represent the sixth rows of each table in which $n_1 = n_2 = 40$ and censoring has percentage composition $40; 20; 40$.

Settings	LR		LW	
	Sad.	Nor.	Sad.	Nor.
Table 4: extreme value	0.12	2.24	0.41	1.66
Table 5: log-logistic	0.08	1.92	0.17	0.69

4. ALGORITHMS FOR COMPUTATION OF NPMLE $\hat{S}_\delta$

Since permutation tests only require a single computation of NPMLE $\hat{S}$, programs, such as the R "interval" package, that compute p-values can base com-

putations on the EM algorithm as long as a check is performed to be sure the EM iterates to the NPMLE and not a local maximum. This is not adequate in the simulation studies of §3.2 wherein each setting required 1000 automated computations of such NPMLEs. Nor is it satisfactory for the test inversion in §3.3 which also required computation of $\hat{S}_\delta$ for hundreds of potential choices of treatment effect δ. In both of these instances, we need to be sure that the algorithm has converged to the NPMLE and both the EM algorithm, as outlined in Peto (1973) and Turnbull (1976), and the hybrid iterative convex minorant (hybrid ICM) in Wellner & Zhan (1997) failed in this automated computation of the NPMLE. The EM algorithm sometimes converged to local maxima, while the hybrid ICM sometimes lead to cumulative sum diagrams that dropped below the x-axis thus resulting in negative probability estimates. For the specific problem we encountered, see Supplementary Material online.

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 to not be the NPMLE, the current survival estimate is used as the input for the hybrid ICM algorithm in Wellner & Zhan (1997) and this algorithm runs until the current estimate is judged to be the NPMLE. To be judged so at either stage in our programs, the current estimate must satisfy the Fenchel conditions given in equation 25 of Wellner & Zhan (1997) with error tolerance $\varepsilon = 10^{-7}$. Our EM-hybrid ICM algorithm succeeded in computing the NPMLE in all our simulations and confidence interval computations. This algorithm underlies all our executable programs for computing permutation significance levels and inverting LR and LW tests and ensures that NPMLEs are used.

5. CONCLUSIONS

We have shown how to use saddlepoint approximations to approximate mid-p-values for treatment benefit in a large class of permutation tests when the data are partly interval-censored data. The approximations are almost always more accurate than the existing normal approximations found in SAS and R software packages.

The speed, stability, and accuracy of these approximations allow us to invert the permutation tests to determine arbitrary $100(1-\alpha)\%$ confidence intervals for the treatment effect under an assumed AFT model. Such intervals have not been previously computed presumably due to the extensive amount of computation. Simulations suggest that the resulting saddlepoint intervals have coverage accuracy which is close to exact. The importance of such test inversion is that it quantifies the treatment benefit. For example, with the breast cosmesis data, rather than simply stating that adjuvant chemotherapy has a statistically significant benefit ($p = 0.0033$ in a one-sided test), the intervals allow for the following confidence statement: "Adjuvant chemotherapy reduces the mean (or median) time to breast retraction by 15.2% to 57.9% with confidence level 95%."

ACKNOWLEDGEMENTS

We thank two anonymous referees and an associate editor for their constructive remarks. This work was supported in part by NSF grant DMS-1104474.

BIBLIOGRAPHY

[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] Butler, R. W. (2007). *Saddlepoint Approximations with Applications*. Cambridge University Press, Cambridge U.K.

[3] Butler, R.W. and Bronson, D.A. (2002). Bootstrapping survival times in stochastic systems by using saddlepoint approximations. *Journal of the Royal Statistical Society series B* 64, 31–49.

[4] Butler, R.W. and Bronson, D.A. (2012). Bootstrap inference in multistate survival models subject to right censoring. *Biometrika*, 99, 959–972.

[5] Davison, A.C. and Hinkley, D.V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* 75, 417–431.

[6] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press, Cambridge U.K.

[7] Fay, M. P. (1996). Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics* 52, 811–822.

[8] Fay, M. P. & Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. *Journal of Statistical Software* 36, 1–34.

[9] Fay, M. P. & Shih, J. H. (2012). Weighted log-rank tests for interval censored data when assessment times depend on treatment. *Statistics in Medicine* 31, 3760–3772.

[10] Feuerverger, A. (1989). On the empirical saddlepoint approximation, *Biometrika* 76, 457–464.

[11] Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* 42, 845–854.

[12] Finkelstein, D. M. & Wolfe, R. A. (1986). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* 41, 933–945.

[13] Fleming, T. & Harrington, D. P. (1981). A class of hypothesis tests for one and two samples censored survival data. *Communications in Statistics* A 10, 763–794.

[14] Gehan, E. A. (1965). A generalized two-sample Wilcoxon test for doubly-censored data. *Biometrika* 52, 650–653.

[15] Hoel, D. G. & Walburg, H. E. (1972). Statistical analysis of survival experiments. *Journal of National Cancer Institute* 49, 361–372.

[16] Lindsey, J. C. & Ryan, L. M. (1998). Tutorial in biostatistics methods for interval-censored data. *Statistics in Medicine* 17, 219–238.

[17] Mantel, N. (1967). Ranking procedures for arbitrarily restricted observation. *Biometrics* 23, 65–78.

[18] Oller, R. & Gómez, G. (2012). A generalized Fleming and Harrington's class of tests for interval-censored data, *The Canadian Journal of Statistics* 40, 501–516.

[19] Oller, R. & Langohr, K. (2013). Tests for right and interval-censored survival data based on the Fleming-Harrington class. http://cran.r-project.org/web/packages/FHtest/index.html.

[20] Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics* 22, 86–91.

[21] Peto, R. & Peto, J. (1972). Asymptotic efficient rank invariant test procedures (with Discussion). *Journal of the Royal Statistical Society, Series A* 135, 185–206.

[22] Prentice, R. L. (1978). Linear rank tests with right-censored data. *Biometrika* 65, 167–179.

[23] 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.

[24] Robinson, J. (1982). Saddlepoint approximations for permutation tests and confidence intervals. *Journal of the Royal Statistical Society series B* 44, 91-101.

[25] 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.

[26] Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions. *Journal of Applied Probability* 24, 875–887.

[27] So, Y., Johnson, G., & Kim, S. H. (2010). Analyzing interval-censored survival data with SAS software. *Proceedings of the SAS Global Forum 2010, Statistics and Data Analysis*, Paper 257-2010, Seattle WA, April 11–14, 2010, http://support.sas.com/resources/papers/proceedings10/257-2010.pdf.

[28] Sun, J. (1996). A non-parametric test for interval-censored data with application to AIDS studies. *Statistics in Medicine* 15, 1387–1395.

[29] Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data.* Springer, New York.

[30] Sun, J., Zhao, Q., & Zhao, X. (2005). Generalized log-rank tests for interval-censored failure time data. *Scandinavian Journal of Statistics* 32, 49–57.

[31] Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* 38, 290–295.

[32] Wellner, J. A. & Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association* 92, 945–959.

[33] Zhao, Q. & Sun, J. (2004). Generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine* 23, 1621–1629.

[34] Zhao, Q. & Sun, J. (2010). Generalized logrank tests for interval-censored failure time data. http://cran.r-project.org/web/packages/glrt/index.html

APPENDIX

Coverage probability error analysis for Table 7. The relative errors in Tables 4 and 5 are not affected by using Monte Carlo mid-p-values in place of exact values as noted in the Supplementary Material online. With 95% confidence interval $[\mathcal{L}, \mathcal{R}]$, let $R\% = 100r\%$ represent the percentage relative error of $\hat{p}(\mathcal{L})$ for $p(\mathcal{L})$ occurring at the 2.5 percentile. We also suppose that $1 - \hat{p}(\mathcal{R})$ has $R\%$ relative error for $1 - p(\mathcal{R})$. Thus we suppose that saddlepoint accuracy is the same in the left and right tails. There are two reasons why such a supposition is reasonable. First, empirical evidence suggests that this is a reasonable assumption when the underlying null log-survival time distribution has a MGF that is convergent in an open neighbourhood of 0 as occurs with our exponential and logistic examples. Secondly, the permutation distribution for U has a central limit theorem so it should be roughly symmetric. Saddlepoint approximations for symmetric distributions preserve tail symmetry (Butler, 2007, §2.1.2) so relative error of saddlepoint approximation is also symmetric. Thus we can expect roughly comparable relative errors at the 2.5 and 97.5 percentiles.

With this supposition, the true values are bounded as

$$0.025(1-r) = \hat{p}(\mathcal{L})(1-r) < p(\mathcal{L}) < \hat{p}(\mathcal{L})(1+r) = 0.025(1+r)$$

$$0.025(1-r) = \{1-\hat{p}(\mathcal{R})\}(1-r) < 1 - p(\mathcal{R}) < \{1-\hat{p}(\mathcal{R})\}(1+r) = 0.025(1+r)$$

so that two-sided coverage error is bounded as

$$0.05 - 0.05r < p(\mathcal{L}) + 1 - p(\mathcal{R}) < 0.05 + .05r.$$

This indicates a relative error of r and an absolute coverage error of $5r\% = (R/20)\%$ as indicated in Table 7.

Received 29 October 2012

Accepted 6 December 2013