Log-rank permutation tests for trend: saddlepoint p-values and survival rate confidence intervals

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

September 10, 2007

Abstract

Suppose $p+1$ experimental groups correspond to increasing dose levels of a treatment and all groups are subject to right censoring. In such instances, permutation tests for trend can be performed based on statistics derived from the weighted log-rank class. This paper uses saddlepoint methods to determine the mid-p-values for such permutation tests for any test statistic in the weighted log-rank class. Permutation simulations are replaced by analytical saddlepoint computations which provide extremely accurate mid-p-values that are exact for most practical purposes and almost always more accurate than normal approximations. The speed of mid-p-value computation allows for the inversion of such tests to determine confidence intervals for the percentage increase in mean (or median) survival time per unit increase in dosage.

Keywords: Mid-p-value; Permutation distribution; Saddlepoint approximation; Trend test; Weighted log-rank class.

1 Introduction

Suppose survival times subject to right censoring are recorded for an experiment with $p+1$ treatment groups. Let the groups correspond to increasing dose levels $l_1 < l_2 < \cdots < l_{p+1}$ in which perhaps dose l_1 is administered as a control. Tests that the group

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

survival functions are stochastically increasing (or decreasing) with dosage level are generally based on the class of weighted log-rank statistics originally suggested in Tarone and Ware (1977). This class includes the log-rank statistic of Mantel (1966), Cox (1972), and Peto and Peto (1972); Gehan's (1965) statistic; the generalized Wilcoxon statistic suggested in Peto and Peto (1972) and Prentice (1978); a specific statistic suggested by Tarone and Ware (1977); and the Fleming and Harrington (1981) class of statistics.

This paper proposes the use of saddlepoint approximations as a means for determining significance levels for tests of trend in the weighted log-rank class under their exact permutation distributions. The speed of such saddlepoint computation also allows these tests to be inverted to yield confidence intervals for the percentage increase in mean (or median) survival time per unit increase in dosage. Such computations would be exceedingly time consuming without the use of saddlepoint methods and consequently no attempts to make such computations have been found in the literature.

Permutation significance was advocated in the original development of such tests by Peto and Peto (1972) however current software such as SAS uses asymptotic normal approximations as described, for example, in Klein and Moeschberger (1997). It will be shown through examples and simulations that saddlepoint approximations are extremely accurate and almost always closer to the true permutation significance levels than the normal approximations.

The computational methods for saddlepoint approximation are exceptionally stable and have been programmed as a general purpose "black box" procedure with executable files available at *http://www.smu.edu/statistics/faculty/butler.html.* In the software, mid-p-values are computed for all five of the weighted log-rank tests mentioned above and exemplified in the paper. In addition, confidence intervals at level 95% are computed for the percentage increase in mean (or median) survival rate by inverting the log rank and generalized Wilcoxon test statistics. Use of the mid-p-value rather than the p-value has been advocated by Agresti (1992), Routledge (1994) and Kim and Agresti (1995) since the ordinary p-value is too conservative. This is particularly so when tests are inverted to determine confidence intervals. Use of ordinary p-values leads to overcoverage, while use of mid-p-values leads to intervals whose attained and nominal coverages are in close agreement.

Previously, double saddlepoint approximations for conditional distributions related to two-sample tests were suggested by Daniels (1958) and later Booth and Butler (1990). The application of such saddlepoint methods to the two-sample log-rank tests was considered in Abd-Elfattah and Butler (2006) and the current paper extends these methods to consider an arbitrary number of treatment levels.

Section 2 provides an overview of the weighted log-rank tests along with the associated permutation distributions that determine their mid-p-values. Saddlepoint approximation to these permutation distributions is addressed in section 3. Section 4 provides numerical examples along with extensive simulations that demonstrate the extraordinary accuracy of the saddlepoint approximations. Section five shows the modifications needed to deal with tied survival times. Section six concludes with confidence interval computation for the percentage increase in survival rate per unit of dosage that is obtained through inversion of these permutation tests.

2 The weighted log-rank class

Suppose that the group sample sizes are $N_1, ..., N_{p+1}$ with a total of N observations. The pooled data are $\{(t_i, z_i, \delta_i) : i = 1, ..., N\}$, where t_i is a time to event, z_i is a $(p+1) \times 1$ treatment indicator, and δ_i indicates that a survival rather than censoring time has been observed. Assume independent censoring with the censoring distribution not dependent on group membership. A test of $H_0 : S_1(t) = \cdots = S_{p+1}(t) = S(t)$, that group survival functions are the same versus the stochastically increasing alternative $H_0 : S_1(t) < \cdots < S_{p+1}(t)$ for all t, is generally based on statistics from the weighted log-rank class.

Let $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ be the distinct ordered survival times among the pooled data with $t_{i1}, \cdots, t_{im_i}$ as the right censored times in the interval $[t_{(i)}, t_{(i+1)})$, $i = 0, 1, \cdots, k$ where $t_{(0)} = -\infty$, $t_{(k+1)} = \infty$ and $k + \sum_{i=1}^{k} m_i = N$. Also, let $z_{(i)}$ and z_{ij} for $i = 1, ..., k$ and $j = 1, ..., m_i$ represent the corresponding $(p+1) \times 1$ indicator vectors for group membership. If we assume no ties among the uncensored data from different groups, then test statistics in the general weighted log-rank class are constructed from $(p+1) \times 1$ vectors of the form

$$v = \sum_{i=1}^{k} w_i \left(z_{(i)} - \frac{1}{n_i} \sum_{l \in R(t_{(i)})} z_l \right), \tag{2.1}$$

where w_i is a weight, n_i is the total number of individuals at risk at time $t_{(i)}^{-}$ and $R(t_{(i)})$ is the set of individuals at risk at $t_{(i)}^{-}$. In tests for trend, the components of v are generally weighted by the increasing dosage levels $l = (l_1, ..., l_{p+1})^T$ with H_0 rejected for small values of $u = l^T v$.

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

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

suggested in Peto and Peto (1972) and Prentice (1978) and referred to as the generalized Wilcoxon, and the general class of tests of Fleming and Harrington (1981) in which the weight function depends on the Kaplan-Meier estimator $\hat{S}(\cdot)$. Here the specific example $w_i = \hat{S}(t_{(i-1)})$ is considered.

In the randomization used for the permutation distribution of v, the survival times and censoring times remain fixed in time order while the $N_1, ..., N_{p+1}$ treatment labels are randomly assigned to the $\binom{N}{N_1, \ldots, N_{p+1}}$ distinct time positions. Saddlepoint approximation for this permutation distribution is simplified by rewriting v in the linear form

$$v = \sum_{i=1}^{k} \left(c_i z_{(i)} + C_i \sum_{j=1}^{m_i} z_{ij} \right) \tag{2.2}$$

where the constants c_i and C_i are fixed constants that depend only on their time position $t_{(i)}$. The weighted log-rank statistic v in (2.1) has a null permutation distribution given as the distribution of (2.2) where $z_{(1)}, \{z_{1j}\}, ..., z_{(k)}, \{z_{kj}\}$ are $(p+1) \times 1$ indicator vectors with uniform distribution over the $\binom{N}{N_1, \ldots, N_{p+1}}$ values for which $\sum_{i=1}^{k}(z_{(i)} + \sum_{j=1}^{m_i} z_{ij}) = (N_1, ..., N_{p+1})$. The weights in (2.2) are

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

3 Saddlepoint approximation for the permutation distribution

The null permutation distribution places a uniform distribution on the set of $(p+1) \times 1$ group indicator vectors $\{z_{(i)}\} \cup \{z_{ij}\}$. This distribution may be constructed from a corresponding set of i.i.d. $p \times 1$ Multinomial $(1, \theta_1, ..., \theta_{p+1})$ indicator vectors which are denoted in capitals by $\{Z_{(i)}^-\} \cup \{Z_{ij}^-\} = \mathbf{Z}^-$. In the reduction from $(p+1)$-dimensional $z_{(i)}$ to p-dimensional $Z_{(i)}^-$, the last component of $z_{(i)}$ has been ignored so $Z_{(i)}^-$ represents the random allocation to the first p components of $z_{(i)}$ with all components of $Z_{(i)}^-$ zero indicating allocation to group $p+1$. The uniform permutation distribution over all one-way designs for which $\sum_{i=1}^{k}(z_{(i)} + \sum_{j=1}^{m_i} z_{ij}) = (N_1, ..., N_{p+1})$ is constructed from the i.i.d. multinomial variates as the conditional distribution of

$$\mathbf{Z}^- = \{Z_{(i)}^-\} \cup \{Z_{ij}^-\} \ \Big| \ \sum_{i=1}^{k} \left(Z_{(i)}^- + \sum_{j=1}^{m_i} Z_{ij}^- \right) = (N_1, ..., N_p)^T = \mathbf{N}_-^T.$$

Writing the trend statistic $u = l^T v$ in terms of the p-dimensional vectors in $\mathbf{Z}^- = \{Z_{(i)}^-\} \cup \{Z_{ij}^-\}$ instead of the $(p+1)$-dimensional $\{z_{(i)}\} \cup \{z_{ij}\}$ and denoting this statistic as $u(Z^-)$, then its permutation distribution is the conditional distribution of the scalar statistic

$$u(\mathbf{Z}^-) \ \mid \ \sum_{i=1}^{k}\left(Z_{(i)}^- + \sum_{j=1}^{m_i} Z_{ij}^-\right) = \mathbf{N}_-^T. \tag{3.1}$$

Simple computations in the Appendix show that

$$u(\mathbf{Z}^-) = l_-^T \sum_{i=1}^{k}\left(c_i Z_{(i)}^- + C_i \sum_{j=1}^{m_i} Z_{ij}^-\right)$$

where $l_- = (l_1 - l_{p+1}, ..., l_p - l_{p+1})^T$.

A saddlepoint approximation for the conditional distribution in (3.1) is constructed in terms of the p-dimensional random variables

$$\begin{aligned} Y &= \sum_{i=1}^{k}\left\{c_i Z_{(i)}^- + C_i \sum_{j=1}^{m_i} Z_{ij}^-\right\} \\ X &= \sum_{i=1}^{k}\left\{Z_{(i)}^- + \sum_{j=1}^{m_i} Z_{ij}^-\right\}. \end{aligned}$$

Assuming any probability vector $\{\theta_1, ..., \theta_{p+1}\}$ for the multinomial distribution, the conditional distribution of $U = l_-^T Y$ given $X = (N_1, ...N_p)^T = \mathbf{N}_-^T$ is the required permutation distribution which can be approximated by using the double saddlepoint approximation of Skovgaard (1987).

Let P be a random variable with the required permutation distribution and let u_0 be the observed value of U. The mid-p-value is $\Pr(P < u_0) + \Pr(P = u_0)/2 =$ mid-$p(u_0)$ and is approximated from the Skovgaard (1987) saddlepoint procedure as the conditional tail probability $\Pr(U \leq u_0 | X = \mathbf{N}_-)$. This approximation uses the joint cumulant generating function for (X, U) given by $K(s,t) = \log M_{X,U}(s,t)$ where

$$M_{X,U}(s,t) = \prod_{i=1}^{k}\left[\left\{\sum_{l=1}^{p}\theta_l \exp(s_l + r_{il}t) + \theta_{p+1}\right\}\left\{\sum_{l=1}^{p}\theta_l \exp(s_l + R_{il}t) + \theta_{p+1}\right\}^{m_i}\right] \tag{3.2}$$

with $s = (s_1, ..., s_p)^T$, $r_{il} = c_{il}(l_l - l_{p+1})$, and $R_{il} = C_{il}(l_l - l_{p+1})$. Then

$$\text{mid-}p(u_0) \simeq \widehat{\Pr}(U \leq u_0 | X = \mathbf{N}_-) = 1 - \Phi(\hat{w}) - \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right), \tag{3.3}$$

where

$$\begin{aligned}\hat{w} &= \text{sgn}(\hat{t})\surd\left(2\left[\left\{K\left(\hat{s}_0, 0\right) - \mathbf{N}_-^T\hat{s}_0\right\} - \left\{K(\hat{s}, \hat{t}) - \mathbf{N}_-^T\hat{s} - u_0\hat{t}\right\}\right]\right) \\ \hat{u} &= \hat{t}\surd\left\{\left|K''(\hat{s}, \hat{t})\right| / |K''_{ss}(\hat{s}_0, 0)|\right\}.\end{aligned}$$

In these expressions, K'' is the $(p+1) \times (p+1)$ Hessian matrix and K''_{ss} is the $p \times p$ sub-block $\partial^2/\partial s \partial s^T$. The numerator saddlepoint $(\hat{s}, \hat{t})$ solves

$$\begin{aligned}K'_{s_l}(\hat{s}, \hat{t}) &= \sum_{i=1}^{k}\left\{\frac{\theta_l \exp(\hat{s}_l + r_{il}\hat{t})}{\sum_{l=1}^{p}\theta_l \exp(\hat{s}_l + r_{il}\hat{t}) + \theta_{p+1}} + \frac{m_i\theta_l \exp(\hat{s}_l + R_{il}\hat{t})}{\sum_{l=1}^{p}\theta_l \exp(\hat{s}_l + R_{il}\hat{t}) + \theta_{p+1}}\right\} = N_l \\ K'_t(\hat{s}, \hat{t}) &= \sum_{i=1}^{k}\left\{\frac{\sum_{l=1}^{p}\theta_l r_{il} \exp(\hat{s}_l + r_{il}\hat{t})}{\sum_{l=1}^{p}\theta_l \exp(\hat{s}_l + r_{il}\hat{t}) + \theta_{p+1}} + \frac{m_i\sum_{l=1}^{p}\theta_l R_{il} \exp(\hat{s}_l + R_{il}\hat{t})}{\sum_{l=1}^{p}\theta_l \exp(\hat{s}_l + R_{il}\hat{t}) + \theta_{p+1}}\right\} = u_0\end{aligned} \tag{3.4}$$

for $l = 1, ..., p$ and the denominator saddlepoint $\hat{s}_0 = (\hat{s}_{10}, ..., \hat{s}_{p0})^T$ solves

$$K'_{s_l}(\hat{s}_0, 0) = \sum_{i=1}^{k}\left\{\frac{\theta_l \exp(\hat{s}_{l0})}{\sum_{l=1}^{p}\theta_l \exp(\hat{s}_{l0}) + \theta_{p+1}} + \frac{m_i\theta_l \exp(\hat{s}_{l0})}{\sum_{l=1}^{p}\theta_l \exp(\hat{s}_{l0}) + \theta_{p+1}}\right\} = N_l \tag{3.5}$$

for $l = 1, ..., p$. Since the computations of $\hat{w}$ and $\hat{u}$ do not depend on the particular value of θ used, the value $\theta_l = N_l/N$ has been used in both (3.4) and (3.5) since it results in an explicit solution for (3.5) as $\hat{s}_0 = 0$ and this simplifies the calculations. For further discussion about this approximation, see Butler (2007, Ch. 4).

The expression in (3.3) uses the saddlepoint approximation as if U, and consequently P, were continuous random variables. The reason that this continuous formula is the appropriate saddlepoint form is that it provides the most accurate approximation for the mid-p-value; see Pierce and Peters (1992), Davison and Wang (2002) and Butler (2007, § 6.1.4) who discuss reasons for this accuracy.

4 Numerical examples and simulations

Two datasets are used to show the accuracy of the saddlepoint methods as compared to normal approximation. The first dataset is the carcinogenicity data of Thomas et al. (1977) which is a small heavily censored dataset with 3 group (censoring) sizes $10(5), 10(4)$, and $9(5)$. Mid-p-values of the 5 weighted log-rank tests for trend are shown in Table 1. "Dosages" in this context are the *decreasing* scores $l = (2,1,0)^T$ so that tests are either rejected for large values of v given in (2.2) or for small values of v using negative scores $-l$.

Table 1. True, saddlepoint, and normal mid-p-values for the log-rank (LR), generalized Wilcoxon (GW), Gehan (GH), Tarone-Ware (TW) and Fleming-Harrington (FH) statistics applied to the two sets of data.

	LR	GW	GH	TW	FH
Thomas et al. (1977) $10(5), 10(4), 9(5)$ $l = (2,1,0)^T$					
True[1] mid-p	0.01725	0.01319	0.00641	0.00981	0.01183
Sadpt.[2] mid-p	0.01693	0.01302	0.00622	0.00974	0.01185
Normal p	0.01697	0.01161	0.00617	0.00895	0.01038
Henderson & Milner (1991) $4(2), 4(2), 4(2)$ $l = (0,1,2)^T$					
True[1] mid-p	0.06504	0.09801	0.12617	0.09112	0.09448
Sadpt.[2] mid-p	0.06393	0.09613	0.12189	0.08819	0.09155
Normal p	0.10214	0.10497	0.12483	0.10197	0.10195

[1] Based on 10^6 simple random samples of $\mathbf{N}_-$ from N and holding the survival/censoring orders fixed. [2] Sadpt. is saddlepoint.

The second dataset is shown in Table 2 and consists of a portion of the data given by Henderson and Milner (1991). The 5 tests for trend using the scores $l = (0,1,2)^T$ are shown in Table 1.

The entries in Table 1 show that saddlepoint approximations are highly accurate for both datasets and consistently more accurate than the normal approximations. Also the saddlepoint method demonstrates considerably greater accuracy when used to approximate mid-p-values for the widely used log-rank (LR) test.

The normal approximations use an asymptotic covariance for v which assumes censoring is independent of the dosage level. This covariance is $\Xi(\sum_{l=1}^{k} w_l^2)$ where

$\mathbf{\Xi} = (\xi_{ij})$ has components

$$\xi_{ij} = \begin{cases} -N_i N_j / N^2 & \text{if} \quad i \neq j \\ N_i(N - N_i)/N^2 & \text{if} \quad i = j. \end{cases} \tag{4.1}$$

This is a generalization to the log-rank class of the covariance expression for the Gehan estimator given in Breslow (1970, pp. 583-4).

The true (simulated) mid-p-values have been calculated by taking 10^6 simple random samples of $\mathbf{N}_-$ from N, holding the censoring orders fixed, and computing the proportion of times that P is less than u_0 plus half the proportion of time it attains u_0.

Table 2. Graft survival times in months of 12 renal transplant patients from Henderson and Milner (1991) with three different levels L0-L2 denoting the total number (0,1,2) of HLA-B or DR antigen mismatches between donor and recipient.

L0	L1	L2
0.068^+	0.101^+	10.66^+
0.508^+	4.410	19.50
13.46	12.21^+	20.90^+
19.73	22.10	32.70

$^+$Right censored.

4.1 Simulation study

Simulation studies were used to show the accuracy of the saddlepoint approximation over a range of data types, numbers of groups, sample sizes, degrees of censoring and error distributions. Two error distributions were used to simulate data and include the log-logistic and Weibull distributions. For each distribution various numbers of groups and group sizes were used. For each consideration, 1000 datasets were drawn from the distribution using a specific censoring percentage and the 1000 saddlepoint and normal p-values were calculated and compared with the 1000 simulated "true" mid-p-values. The censored data were selected at random, independently of the data generation, and before allocation of the data to the various groups. In the group allocation, N_i values were assigned to group i and the log-survival times were shifted in location by the amount $i\beta$, where the value of β was chosen to approximately achieve a p-value of 5%. The doses were chosen to be the equally spaced values $\{0, 1, 2, ...\}$. The simulated mid-p-values associated with each consideration were computed as follows. For each of the 1000 datasets, 10^6 permutations of the test statistic were computed by holding the survival/censoring positions fixed. The simulated mid-p-value is then the proportion of such generations that are less than the observed statistic plus half the proportion that are equal. These calculations were implemented for both log-rank and generalized Wilcoxon type trend tests. Tables 3 and 4 show the results for the two distributions respectively.

Each table provides the following information: the "Mean" is the average true mid-p-value (based on 10^6 simulations) over the 1000 datasets, "Sadpt. Prop." is the proportion of the 1000 datasets for which the saddlepoint mid-p-value was closer to the true mid-p-value than the normal p-value, "Abs. Err. Sadpt." is the average absolute error of the saddlepoint mid-p-value from the true mid-p-value, "Rel. Abs. Err. Sadpt." is the average relative absolute error of the saddlepoint mid-p-value from the true mid-p-value, and the remaining listings are the same assessments for the normal approximation using the covariance estimate in (4.1).

For the log-rank simulations, the saddlepoint mid-p-value was more accurate in 98.3% of the overall cases as compared to the normal approximation. For the generalized Wilcoxon simulation, the saddlepoint was only slightly worse achieving greater accuracy in 89.7% of the overall cases. In both tables, the average absolute saddlepoint error was less than 0.001 with average relative error typically less than 0.01%.

Additional simulations that have not been reported considered the other three tests (GH, TW, and FH) as well as a variety of other data types that reflect varying amounts of censoring, small and large sample sizes, and varying degrees of imbalance in dosage allocation. With all five tests and under all conditions the saddlepoint approximations were found to be highly accurate and generally superior to the normal approximations in replicating the exact permutation significance.

Table 3. Performance under simulation from the log-logistic distribution. Notation 0^m signifies m repetitions of 0 so that $0.0^m nop = n.op \times 10^{-m-1}$.

Stat.	Mean	Sadpt.	Abs. Err.	Abs. Err.	Rel. Abs.	Rel. Abs.
		Prop.	Sadpt.	Normal	Err. Sadpt.	Err. Nor.
3 groups, sizes $= 7, 8, 9.$				$\beta = 1$	30% censoring	
LR	0.044	0.973	0.0^3364	0.0^2708	0.0^5666	0.0^5807
GW	0.035	0.969	0.0^3131	0.0^2130	0.0^6406	0.0^5828
3 groups, sizes $= 5, 15, 25.$				$\beta = 1$	15% censoring	
LR	0.030	0.997	0.0^3173	0.0^2661	0.0^5602	0.0^3118
GW	0.019	0.954	0.0^3127	0.0^3624	0.0^5261	0.0^5362
3 groups, sizes $= 25, 20, 30.$				$\beta = 0.225$	30% censoring	
LR	0.071	0.977	0.0^3194	0.0^2306	0.0^4111	0.0^3527
GW	0.077	0.867	0.0^3151	0.0^3421	0.0^5413	0.0^4281
5 groups, sizes $= 7, 8, 7, 7, 8.$				$\beta = 0.43$	15% censoring	
LR	0.059	0.994	0.0^3285	0.0^2463	0.0^4114	0.0^2209
GW	0.045	0.931	0.0^3136	0.0^3808	0.0^4284	0.0^3533
5 groups, sizes $= 15, 12, 15, 10, 13.$				$\beta = 0.15$	30% censoring	
LR	0.138	0.975	0.0^3256	0.0^2314	0.0^4233	0.0^3178
GW	0.130	0.810	0.0^3214	0.0^3512	0.0^5425	0.0^4233

Table 4. Performance under simulation from the Weibull distribution.

Stat.	Mean	Sadpt.	Abs. Err.	Abs. Err.	Rel. Abs.	Rel. Abs.
		Prop.	Sadpt.	Normal	Err. Sadpt.	Err. Nor.
3 groups, sizes $= 9, 7, 8.$				$\beta = 0.65$	5% censoring	
LR	0.030	0.990	0.0^3332	0.0^2581	0.0^3198	0.0149
GW	0.044	0.942	0.0^3267	0.0^2133	0.0^3132	0.0^2205
3 groups, sizes $= 15, 5, 25.$				$\beta = 0.15$	15% censoring	
LR	0.147	0.946	0.0^3678	0.0^2648	0.0^5331	0.0^4168
GW	0.166	0.790	0.0^3506	0.0^3852	0.0^6480	0.0^6343
5 groups, sizes $= 7, 8, 7, 7, 8$				$\beta = 0.2$	30% censoring	
LR	0.070	0.986	0.0^3295	0.0^2481	0.0^4104	0.0^4616
GW	0.091	0.920	0.0^3188	0.0^3900	0.0^6838	0.0^5609
5 groups, sizes $= 5, 15, 15, 5, 10.$				$\beta = 0.2$	30% censoring	
LR	0.049	1.00	0.0^3149	0.0^2245	0.0^5641	0.0^4990
GW	0.074	0.907	0.0^3153	0.0^3712	0.0^5535	0.0^4168
5 groups, sizes $= 15, 12, 15, 10, 13.$				$\beta = 0.15$	30% censoring	
LR	0.055	0.998	0.0^3156	0.0^2266	0.0^5864	0.0^4534
GW	0.082	0.882	0.0^3159	0.0^3524	0.0^5396	0.0^4134

5 Tied survival times

Suppose there are $d_i \geq 1$ tied survival times at epoch $t_{(i)}$ for $i = 1, ..., k$. Simple computations show that the expression for v in (2.2) can be used with each of the tied survivals at $t_{(i)}$ using weight c_i and each of the censored values in $[t_{(i)}, t_{(i+1)})$ using weight C_i given by

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

Thus, with the appropriate assignment of scores, the permutation distribution of u can be approximated by using the Skovgaard expression as in §3.

As an example, consider the accelerated life tests on electrical insulation by Schmee and Hahn (1979) for $p = 4$ groups of heat levels $\{150°, 170°, 190°, 220° C\}$. Half their data have been used and are given in Table 5.

Table 5. Accelerated life test data taken from Schmee and Hahn (1979) with $p = 4$ groups of heat levels.

$220° C$	$190° C$	$170° C$	$150° C$
408	408	1764	8064^+
504	1344	3542	8064^+
528^+	1344	4860	8064^+
528^+	1440	5448^+	8064^+
528^+	1680^+	5448^+	8064^+

$^+$Right censored.

Dosage weights were taken as $l = (0, 1, 2, 3)^T$. Since smaller failure times are anticipated in higher heat groups, the trend test rejects for large values of u. Table 6 compares exact mid-p-values, determined by simulating 10^6 permutations of u, with saddlepoint and normal mid-p-values.

Table 6. The accelerated life test data using $l = (0, 1, 2, 3)^T$. Table entries are as described in Table 1.

	LR	GW	GH	TW	FH
True[1] mid-p	.00279	.00194	.00183	.00196	.00194
Sadpt. mid-p	.00265	.00189	.00182	.00196	.00192
Normal p	.00480	.00394	.00410	.00400	.00402

6 Confidence interval for percentage increase in mean (or median) survival time per unit dosage.

Prentice (1978) formulated tests for trend as tests for the dosage slope in a log-linear rank model subject to right censoring. If T is an uncensored survival time, then $\log T$ is assumed to have location parameter $\mu + l_i \beta$ for the ith treatment group with dosage l_i and all groups are assumed to share a common error distribution. Let the unordered log-survival/censored times be denoted by the N-vector $y = (\log t_1, ..., \log t_N)^T$ with $x = (x_1, ..., x_N)^T$ indicating the dosage levels for corresponding components that assume values from $\{l_1 < \cdots < l_{p+1}\}$. The framework of the censored accelerated failure time model, as described in Kalbfleisch & Prentice (2002), determines the confidence interval for β, the log of the increase in mean survival per unit of dosage. While the rank tests of §3 were concerned with testing $H_0 : \beta = 0$ vs. $H_1 : \beta > 0$ essentially using the components of y, these same tests provide for testing $H_0 : \beta = \beta_0$ vs. $H_1 : \beta > \beta_0$ if the log-survival/censored time residuals $y - x\beta_0$ are used in place of y. Within this framework, a 95% confidence interval consists of those β_0 values whose mid-p-values in (3.3) fall within the range $[0.025, 0.975]$.

The 95% confidence interval for β on the log-time-scale is more meaningfully reported as a 95% confidence interval on the time scale as $100(e^\beta - 1)$, the percentage increase in mean (or median) survival time per unit dosage. To understand this interpretation, suppose that T_i is a survival time using dosage i in the log-linear rank model. Then a unit increase in dosage leads to the percentage increase in mean (or median) survival times as

$$100 \left\{ \frac{E(T_{i+1})}{E(T_i)} - 1 \right\} = 100 \left\{ \frac{e^{\mu+(i+1)\beta} E(e^\varepsilon)}{e^{\mu+i\beta} E(e^\varepsilon)} - 1 \right\} = 100 \left(e^\beta - 1 \right). \tag{6.1}$$

Confidence intervals for (6.1) can be computed by using the executable files that are available along with instructions at *http://www.smu.edu/statistics/faculty/butler.html.*

The datasets from Table 1 have been used to construct confidence intervals that are given in Table 7. The true, saddlepoint, and normal confidence intervals are given by inverting their corresponding test procedures. The Thomas data set used dosages $2, 1$, and 0 so one-sided tests for $H_0 : \beta = \beta_0$ vs. $H_0 : \beta > \beta_0$ were calculated from $\log t - x\beta_0$ where x as defined above. Exact, saddlepoint, and normal mid-p-values were computed using incremental steps of $\Delta\beta_0 = 0.001$. Since plots of the true, saddlepoint, and normal mid-p-values vs. β_0 are step functions which cannot exactly attain the end values 0.025 and 0.975, conservative intervals are reported in Table 7 using end values of β_0 that take the first step below 0.025 and above 0.975.

For the Henderson and Milner data, the normal approximation for permutation significance of the log rank test was not able to step above 0.975 when calculated using large values of β_0 therefore the upper range of the confidence interval has been reported as ∞. Note that the three methods for inverting the generalized Wilcoxon test lead to the same 95% confidence intervals. This happens because all three mid-p-values jump below 0.025 and above 0.975 at the same β_0 values but, in doing so, have used quite different mid-p-values.

Table 7. Confidence intervals for the percentage increase in mean (or median) lifetime per unit of dosage.

	LR		GW	
	lower	upper	lower	upper
Thomas et al. (1977) $10(5), 10(4), 9(5)$				
True	-51.471	-9.6067	-46.794	-10.506
Sadpt.	-51.471	-9.6067	-46.794	-10.506
Normal	-57.387	-8.1488	-46.794	-9.6067
Henderson & Milner (1991) $4(2), 4(2), 4(2)$				
True	-11.041	641.87	-11.839	641.87
Sadpt.	-11.041	641.87	-11.839	641.87
Normal	-77.665	∞	-11.839	641.87

References

[1] Abd-Elfattah, E.F. and Butler, R.W. (2007). The weighted log-rank class of permutation tests: p-values and confidence intervals using saddlepoint approximations. To appear *Biometrika.* Technical Report available at *http://www.smu.edu/statistics/TechReports/tech-rpts.asp*

[2] Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, **7**, 131-153.

[3] Booth, J.G. and Butler, R.W. (1990). Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika* **77**, 787-796.

[4] Breslow, N. (1970). A generalized Kruskal-Wallis type test for comparing K samples subject to unequal patterns of censorship. *Biometrika* **57**, 579-594.

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

[6] Cox, D.R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society, series B* **20**, 215-242.

[7] Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, series B* **34**, 187-220.

[8] Daniels, H. E. (1958). In discussion of Cox (1958), 236-238.

[9] Davison, A.C. and Wang, S. (2002). Saddlepoint approximations as smoothers. *Biometrika* **89**, 933-938.

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

[11] Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203-223.

[12] Henderson, R. and Milner, A. (1991). Aalen plots under proportional hazards. *Applied Statistics* **40**, 401-409.

[13] Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data.* 2nd Ed. New York: Wiley.

[14] Kim, D. and Agresti, A. (1995). Improved exact inference about conditional association in three-way contingency tables. *Journal of the American Statistical Association* **90**, 632-639.

[15] Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis, Techniques for Censored and Truncated Data.* New York: Springer-Verlag.

[16] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163-170.

[17] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, series A* **135**, 185-206.

[18] Pierce, D.A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *Journal of the Royal Statistical Society, series B* **54**, 701-737.

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

[20] Routledge, R.D. (1994). Practicing safe statistics with the mid-*p*. *Canadian Journal of Statistics* **22**, 103-110.

[21] Schmee, J. and Hahn, G.J. (1979). A simple method for regression analysis with censored data. *Technometrics* **21**, 417-432.

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

[23] Tarone, R. and Ware J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156-160.

[24] Thomas D. G., Breslow N. E. and Gart J. J. (1977). Trend and homogeneity analyses of proportions and life table data. *Computers and Biomedical Research*, **10**, 373-381.

7 Appendix

In order to write $u(Z)$ as $u(Z^-)$, first denote $Z_{(i)} = (Z_{(i)1}, ..., Z_{(i),p+1})^T = (\{Z^-_{(i)}\}^T, Z_{(i),p+1})^T$ and $Z_{ij} = (Z_{ij,1}, ..., Z_{ij,p+1})^T = (\{Z^-_{ij}\}^T, Z_{ij,p+1})^T$. Then

$$\begin{aligned} u &= l^T \sum_{i=1}^{k} \left\{ c_i Z_{(i)} + C_i \sum_{j=1}^{m_i} Z_{ij} \right\} \\ &= \sum_{i=1}^{k} \left[l_p^T \left(c_i Z^-_{(i)} + C_i \sum_{j=1}^{m_i} Z^-_{ij} \right) + l_{p+1} \left(c_i Z_{(i),p+1} + C_i \sum_{j=1}^{m_i} Z_{ij,p+1} \right) \right]. \end{aligned}$$

where $l_p^T = (l_1, ..., l_p)$. Since $Z_{(i),p+1} = 1 - \mathbf{1}^T Z^-_{(i)}$ and $Z_{ij,p+1} = 1 - \mathbf{1}^T Z^-_{ij}$ with $\mathbf{1} = (1, ...1)^T$ as $(p \times 1)$, then

$$\begin{aligned} u &= \sum_{i=1}^{k} \left[c_i l_p^T Z^-_{(i)} + C_i \sum_{j=1}^{m_i} l_p^T Z^-_{ij} + l_{p+1} \left\{ c_i (1 - \mathbf{1}^T Z^-_{(i)}) + C_i \sum_{j=1}^{m_i} (1 - \mathbf{1}^T Z^-_{ij}) \right\} \right] \\ &= \sum_{i=1}^{k} \left\{ c_i (l_p^T - l_{p+1} \mathbf{1}^T) Z^-_{(i)} + C_i \sum_{j=1}^{m_i} (l_p^T - l_{p+1} \mathbf{1}^T) Z^-_{ij} \right\} + Q \\ &= l_-^T \sum_{i=1}^{k} \left\{ c_i Z^-_{(i)} + C_i \sum_{j=1}^{m_i} Z^-_{ij} \right\} + Q, \end{aligned}$$

where $Q = l_{p+1}(\sum_{i=1}^{n} c_i + m_i C_i) = 0$; see Kalbfleisch and Prentice (2002, eqn. 7.20) for the details about why $Q = 0$.