Tests for Symmetry with Right Censoring

Ehab F. Abd-Elfattah
Department of Mathematics, Faculty of Education
Ain Shams University, Cairo, Egypt

and

Ronald W. Butler*
Department of Statistical Science, Southern Methodist University
Dallas, Texas, 75275, U.S.A.

25 April 2009

Abstract

Permutation tests for symmetry are suggested for data that are subject to right censoring. Such tests are directly relevant to the assumptions that underlie the generalized Wilcoxon test since the symmetric logistic distribution for log-errors has been used to motivate Wilcoxon scores in the censored accelerated failure time model. Its principal competitor is the log-rank test motivated by an extreme value error distribution that is positively skewed. The proposed one-sided tests for symmetry against the alternative of positive skewness are directly relevant to the choice between usage of these two tests.

Keywords: Censoring; Mid-p-value; permutation distribution; saddlepoint approximation; symmetry; weighted log-rank class

AMS Subject Classification: 62G10 Primary; 62N01 Secondary.

1 Introduction

Tests of symmetry for a distribution of log-survival times are proposed when the data are subject to independent right censoring. Such tests are relevant in the censored accelerated failure time (AFT) model ([10], ch. 7) where log-logistic errors have traditionally been used to motivate the scores for the generalized Wilcoxon test. Given the need and relevance in checking for such symmetry, we have not found any such tests in the literature that can deal with the additional complication of right

* *Corresponding author*: Email: rbutler@smu.edu

censoring. This paper shows how such tests can be performed by using two-sample weighted log-rank tests that are commonly used in survival analysis.

Initially suppose the median is known (a condition to be removed) and there is no censoring (also to be removed). Let data consist of unordered log-survival times $y_i = \log t_i$ for $i = 1, ..., N$ which form a random sample from continuous distribution G with known median 0. In this context, a test for symmetry of G about 0 by using the two-sample Wilcoxon/Mann-Whitney test is suggested by Gupta in [7]. The data are "folded" about zero and taken to be $|y_1|, ..., |y_N|$. Indices $\{i : y_i > 0\}$ in the right tail are designated as the "treatment" group and indices $\{i : y_i < 0\}$ in the left tail comprise the "control" group. If the null holds and $Y \sim G$ is symmetric, then the folded left tail is the same as the right tail; e.g. the conditional distributions $-Y \,|\, Y < 0$ and $Y \,|\, Y > 0$ are the same so that $|y_1|, ..., |y_N|$ are a random sample from a common distribution. In [20] this two-sided test is shown to be at least as powerful as the test of McWilliams in [13], who in turn shows his test to be more powerful than tests by Butler in [2], Rothman and Woodroofe in [18], and Hill and Rao in [8] for selected alternatives in the asymmetric lambda distribution class.

Now let the median be unknown and suppose data are uncensored. The point of symmetry is the median of G whose estimate $\hat{m}$ is subtracted as described in [9], §3.9. The test is now based upon testing symmetry about zero using centered data $\{y_i - \hat{m}\}$.

The current paper supposes the median of G is unknown but includes the added complication that data are independently right censored. Now centering must subtract a median estimate that accounts for the censoring and so the median estimate determined by inverting the Kaplan-Meier estimator is used as described in section 2.1 below. Let $\{y_i^*\}$ be the unordered log-survival/censoring times that have been centered by using the Kaplan-Meier estimate for median. Treatment labels are assigned to $\{i : y_i^* > 0\}$, the right tail, and control labels are assigned to $\{i : y_i^* < 0\}$, the left tail. Values of $\{y_i^*\}$ are now folded about 0, however the folded left-tail values

are not simply $\{|y_i^*|\}$ when y_i^* results from a censoring time; this leads to a substantial and unavoidable loss of information as discussed in section 2.2. Now, in this two sample context, the folded values can be used to compute a weighted log-rank statistic v that accounts for the censoring. A sufficiently small (large) v indicates that the treatment distribution or right tail is stochastically larger (smaller) than its control counterpart, the left tail, so that Y is positively (negatively) skewed.

The test that rejects the hypothesis of symmetry of Y for small v is the one-sided test used with the alternative hypothesis that Y is positively skewed. Symmetry assumptions for Y underlie the generalized Wilcoxon weights for v whereas positively skewed assumptions for Y are consistent with Y having an extreme value distribution that would motivate log-rank weights. Thus the use of this one-sided test helps to distinguish logistic versus extreme value errors for homogeneous groups in the AFT model as discussed in section 2.2. The mid-p-value associated with this test measures the degree of skewness ranging from positive skewness (small mid-p-value) to negative skewness (large mid-p-value).

Numerical examples of such one-sided tests are given in section 3, and simulated level and power for these tests are described in section 4.

The simulations suggest that the symmetry test from the log-rank class that uses log-rank weights maintains its level well with light censoring and becomes conservative under very heavy censoring by not rejecting often enough. By contrast, the test using generalized Wilcoxon weights is the opposite; it is too liberal under light censoring and has accurate level with very heavy censoring. In power simulations, tests using the log-rank weights demonstrated greater power for the most part with all levels of censoring up to 50%. In these power simulations, the alternative was that Y has a positively skewed extreme value distribution. Our focus on this particular alternative is based on the need to decide between the use of Wilcoxon or log-rank weights when picking v as a member of the log-rank class.

A consequence of right censoring is that weighted log-rank tests end up being

computed but using only a subset of the centered data $y_1^*, ..., y_N^*$. This subset consists of all of the centered data from the right tail along with only the centered log-survival times from the left tail. Censored data from the left tail are not involved in test statistic computation. This oddity might naturally lead to the impression that censored control data are not informative about the alternative under test. However, as discussed in section 2.2, these data do determine the median estimate that defines the control and treatment groups so they are somewhat informative. However, once the groups are determined, the left-tail sample size is diminished since left-tail values are not informative about the left tail of G. Consequently, this severely affects the power in testing for symmetry. Since there is no alternative but to fold the left tail onto the right tail for comparison, there appears nothing that can be done about the loss of information and reduction in power. One must accept that there is no further information available from the data for testing symmetry.

2 Test construction

2.1 *Centering and folding data that have been censored*

Testing for symmetry inevitably entails choosing a folding point as the hypothesized point of symmetry, folding the left tail onto the right (or equivalently the right onto the left), and making comparisons of the folded distributional shapes. This folding point is $S^{-1}(0.5)$, the median of log-survival function $S(t)$ which, because of censoring, is chosen as the median $\hat{S}^{-1}(0.5)$ from the Kaplan-Meier estimate $\hat{S}$. Estimate $\hat{S}$ uses the log-survival/censored data $\{y_i, \delta_i : i = 1, \ldots, N\}$, with δ_i as a survival indicator of y_i, and is a decreasing step function with steps located at log-survival points. Thus the value $\hat{S}^{-1}(0.5)$ is either a single log-survival time or the range of values in between two log-survival times should there be a horizontal step at height 0.5. The former setting is simpler and more common and is discussed in detail below. In the latter setting, $\hat{S}^{-1}(0.5)$ is taken as the midpoint of this horizontal step and this case

is discussed in section 2.4.

In the setting for which the median estimate is a single log-survival time, the centered median becomes 0 which, for purposes of testing symmetry, should not be considered as belonging to either the treatment or control group. The value $y^* = 0$ has no discriminatory value between the two groups and therefore should not enter into the computation of the test statistic for symmetry.

The small data set given in the left-most column of Table 1 illustrates the centering and folding of the data in order to test for symmetry. The median estimate is the unique value $\hat{S}^{-1}(0.5) = 4$ so responses $1, 2^+$, and 3 form the control group and 5^+ and 6 are treatment group. Centered survival times are $y_i^* = y_i - \hat{S}^{-1}(0.5)$, however the centering of censored data is best understood in terms of centering the survival range. For example, survival range $2^+ = (2, \infty)$ is centered to range $-2^+ = (-2, \infty)$ or more generally y_i^+ is centered to (y_i^*, ∞).

Folding about zero is straightforward for all survival times and for censored values in the right tail. Centered treatment range $(1, \infty)$ folds onto $1^+ = (1, \infty)$ and is unchanged. However, all censored values in the left tail fold into $[0, \infty)$ as, for example, occurs with $(-2, \infty)$.

For the purpose of weighted log-rank test construction, folded survival times $\{y_i^*\}$, and ranges $\{(y_i^*, \infty)\}$ can be described by the triple (f_i, z_i, δ_i). Here, f_i is either $|y_i^*|$ or the left edge of the folded range, z_i is the group indicator, and δ_i indicates right censored. From the example, one can see that censored control values in the left tail always fold into the triple $(0, 0, 0)$ and censored treatment values always fold into $\{y_i - \hat{S}^{-1}(0.5), 1, 0\}$.

2.2 *One-sided log-rank tests*

Weighted log-rank tests can now be applied to the pooled folded data $\{f_i, z_i, \delta_i : i = 1, \ldots, N-1\}$. Suppose $0 < f_{(1)} \leq f_{(2)} \leq \cdots \leq f_{(k)}$ are the ordered values from $\{f_i : \delta_i = 1\}$, the log-survival values that have been centered and folded (excluding

the folded median for which $f_i = 0$). The weighted log-rank test statistic that tests symmetry with weights $\{w_i : i = 1, \ldots, k\}$ is

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

where $z_{(i)}$ is a treatment indicator for $f_{(i)}$, and n_i is the size of the risk set at $f_{(i)}$, or $R\{f_{(i)}\}$.

Computation of v for the example in Table 1 uses the last four columns to give

$$v = w_1 \left\{ 0 - \frac{2}{4} \right\} + w_2 \left\{ 1 - \frac{1}{2} \right\} + w_3 \left\{ 0 - \frac{1}{1} \right\}.$$

The risk set $R\{f_{(1)}\} = R(1)$ includes 4 out of the 6 data points. The two data points notably missing are the censored control value which folds over to 0 and the median estimate which folds to 0. Neither of these values is in $R\{f_{(1)}\}$ or $R\{f_{(i)}\}$ for $i > 1$.

The simple example illustrates two important general points about the data points that do not enter into the computation of v in (1). Censored data from the left tail fold into the value $(0, 0, 0)$ and do not enter into the computation of v. Neither does the data point associated with the median. Thus all censored values in the left tail are "uninformative" after folding since they do not enter into log-rank test statistic computation. Of course censored values in the left tail are informative in the sense that their presence in the data contribute to determining the data center through value $\hat{S}^{-1}(0.5)$.

The log-rank test rejects for small values of v which suggests that the distribution for the right tail is stochastically larger than the distribution for the folded left tail. Such a situation generally occurs when the log-survival distribution is positively skewed as given in the following result that can be easily shown.

Lemma 1 *Suppose $m = G^{-1}(0.5)$ is the unique median of G. The distribution for the control group is $1-2G(m-y)$ for $y \in (0, \infty)$ while the distribution for treatment is $2G(m+y)-1$ for $y \in (0, \infty)$. When Y has a symmetric distribution, these distributions*

are the same, e.g.

$$1 - 2G(m - y) = 2G(m + y) - 1 \qquad \forall y > 0. \tag{2}$$

Under the alternative in which the treatment CDF is stochastically larger, the equality in (2) is replaced by $\geq$ *with strict* $>$ *for some* y.

The two cases of particular importance are when G assumes a logistic distribution and an extreme value distribution with CDF $\exp(-e^{-y})$. These two error distributions in the AFT model motivate score tests whose weights in the log-rank class of tests are the generalized Wilcoxon and log-rank weights commonly used in survival analysis. These two distributions may be distinguished by the fact that equality (2) holds for the symmetric logistic and strict inequality occurs with an extreme value distribution for all $y > 0$ as formalized below. Thus these two distributional classes are in the null and alternative hypotheses respectively for the one-sided test described.

Lemma 2 *If* G *is an extreme value distribution, then the relationship in (2) is strict for all* $y > 0$ *with equality holding only for* $y = 0$.

Proof. For this distribution, the relationship in (2) can be shown equivalent to

$$2^{-e^{y}} + 2^{-e^{-y}} \leq 1.$$

This inequality is strict for $y \neq 0$ and equal only for $y = 0$. ■

2.3 *Negative skewness*

A test for symmetry against the alternative that the distribution of Y is negatively skewed entails testing that $-Y$ is positively skewed. Such tests start with negative data $-y_1, ..., -y_N$ which get centered and folded. Using the explanation given in the last paragraph, it can be shown that the folded data for testing negative skewness are exactly the same data used for testing positive skewness except that the

treatment/control labels are reversed. This leads to folded data $\{f_i, 1 - z_i, \delta_i : i = 1, \ldots, N-1\}$. The resulting weighted log-rank test statistic v_- in (1) is

$$v_- = \sum_{i=1}^{k} w_i \left(1 - z_{(i)} - \frac{1}{n_i} \sum_{l \in R\{f_{(i)}\}} (1 - z_l) \right) = \sum_{i=1}^{k} (w_i - 1) - v_+$$

where v_+ is used for testing positive skewness.

This shows that rejecting for small v_- when testing for negative skewness is equivalent to rejecting for large v_+ using the test statistic for positive skewness. Thus the mid-p-value of the one-side test that rejects for small v_+ is a measure of the degree of skewness with small (large) values indicating positive (negative) skewness and target 0.5 representing symmetry.

Several points must be made to understand why negative data fold into $\{f_i, 1 - z_i, \delta_i : i = 1, \ldots, N-1\}$. First, the division into treatment/control groups for the negative data is the same as for the positive data with the group labels reversed. This occurs because the Kaplan-Meyer estimates for the two data sets are related by $\hat{S}_{-Y}^{-1}(0.5) = -\hat{S}_Y^{-1}(0.5)$, e.g. the median estimate for negative data is the negative of the median estimate for the original data. Secondly, the negative data are now left censored. As a result, negative control data that are left censored do not fold into $[0, \infty)$ but rather into the right censored regions (y_i^*, ∞) of the treatment group from the positive data. These right censored regions that come from the negative control data group now enter into the computation of v_-. Thirdly, negative treatment data that are left censored fold into $[0, \infty)$, the same folded regions of the positive control data that are right censored.

2.4 *Median is not unique*

If $\hat{S}^{-1}(0.5)$ assumes a range of values $[A, B]$ with

$$A = \inf\{y : \hat{S}(y) = 0.5\}, \qquad B = \sup\{y : \hat{S}(y) = 0.5\},$$

then $\hat{S}^{-1}(0.5) = (A + B)/2$ is used for centering. With this choice, all censored values within the range (A, B) end up not in the risk set $R(f_{(1)})$ and therefore do

not enter into computation of v. This is because those in the left half of (A, B) are censored controls, while those on the right half lie in $[0, f_{(1)})$ and thus are also not in $R(f_{(1)})$. Thus all censored log-times falling within the median range (A, B) do not affect computation of v.

When $\hat{S}^{-1}(0.5)$ is chosen to be $(A+B)/2$, a complication also arises that involves tied absolute log-survival times. Subtracting the midpoint of $[A, B]$ from the log-survival times A (control) and B (treatment) results in folded centered log-survival times that are tied at $(B-A)/2$ and from opposite groups; e.g. $\{(B-A)/2, 0, 1\}$ for A and $\{(B-A)/2, 1, 1\}$ for B. The recommendation in [10], p. 234 for dealing with such a tie is to average the two mid-p-values determined from the two configurations that are consistent with the tie. Thus, in this instance, our computation of overall mid-p-value averages the two mid-p-values that result from permuting the treatment/control labels for the tie at $(B - A)/2$.

A median choice other than $\hat{S}^{-1}(0.5) = (A + B)/2$ would give unequal treatment to values A and B in the computation of v. This lack of balanced treatment seems unnatural and, in fact, is a good reason for selecting $(A+B)/2$ as the median estimate.

3 Examples

Our examples consider exact permutation tests for symmetry as originally recommended in [14]. One-sided permutation mid-p-values for weighted log-rank statistics are computed by fixing the set of pairs $\{(f_i, *, \delta_i : i = 1, \ldots, N-1\}$ and permuting the labels for treatment and control assigned to $*$ so as to maintain a fixed value of $z_\bullet = \sum_{i=1}^{N-1} z_i$. "Exact" mid-$p$-values have been computed by simulating 10^6 values of v from all of its $\binom{N-1}{z_\bullet}$ possible values and then computing the permutation significance for the observed value of v. Saddlepoint approximations for the exact mid-p-values are also computed that require no simulation. These approximations almost exactly reproduce the exact permutation mid-p-values and have been described

in [1]. The third approximation is the standard normal approximation used in SAS and described, for example, in [11].

Three examples are used to illustrate these tests for symmetry. Data in the first example have no censored values while data in the last two examples are subject to heavy censoring.

Example 1. Uncensored data in [6] measure the percentage of silica for $N = 22$ chondrites meteors. The data are to be tested for symmetry about the median value $\{\log(28.69) + \log(29.36)\}/2$ which leads to 11 in both groups. Using the Wilcoxon signed rank test, the "exact" permutation mid-p-value from simulation is 0.36156, the saddlepoint mid-p-value is 0.36166, and the normal approximation yields 0.42914.

Example 2. Table 2 shows 22 survival/censoring times, in months, taken from [15] for patients suffering from chronic active hepatitis treated with prednisolone.
Median estimate $\hat{S}^{-1}(.5) = 4.98$ is a log-survival time and seven censored values to its left do not enter into test computation. Note that the treatment value 5.00^+ is retained since after centering it assumes the value 0.02^+ just to the right of the smallest absolute centered log-survival time $f_{(1)} = 0.02$.

Table 3 provides mid-p-values for 5 tests from the weighted log-rank class. These tests included the log-rank (LGR) test ($w_i \equiv 1$), the generalized Wilcoxon (GWL) test

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

advocated by Peto and Peto [14], and Prentice [17], Gehan's [5] (GH) test ($w_i = n_i$), the Tarone-Ware [19] (TW) test ($w_i = \sqrt{n_i}$), and a test (FH) in the Fleming and Harrington [4] class in which $w_i = \hat{S}_F(f_{(i-1)})$ where $\hat{S}_F$ is the Kaplan-Meier estimator of the folded data. Note the lack of significance for the assumption of symmetry which may be attributed to the lack of information about the right tail in which there is only one survival and 4 censoring times.

A comparison of the exact mid-p-values in Table 3 with their saddlepoint (Sadpt.) approximations demonstrates the very striking accuracy of the latter and contrasts with the poor performance of the normal approximation.

Example 3. Censored survival times for patients with non-Hodgkins lymphoma are taken from [3]. The asymptomatic portion of the log-times are used and are displayed in Table 4. The value $\hat{S}^{-1}(0.5)$ coincides with the flat inter-survival step $(5.684, 5.707)$ so midpoint 5.6955 is used for centering and the censored log-time 5.704^+ is uninformative. After centering and folding, the control response at 5.684 and treatment response 5.707 lead to a tie at the value 0.0115. This leads to 15 control survivals and 14 treatment responses among which 12 are censored. As a result of the tie, the mid-p-values in Table 5 have been computed as the average of the two mid-p-values that are consistent with the tie.

4 Simulation of level and power for one-sided tests

4.1 *Level*

Q-Q plots were constructed to determine if the distribution of mid-p-values for the one-sided tests are uniform under the null hypothesis of symmetry. In the weighted log-rank class of tests, only the more commonly used tests with log-rank and generalized Wilcoxon weights were considered.

Suppose that log-survival has the logistic distribution $G(y) = (1 + e^{-y})^{-1}$ that is symmetric about 0 and log-censoring has the log-Weibull $(2, 9)$ distribution with Weibull (a, b) density

$$f(t; a, b) = a/b\,(t/b)^{a-1} \exp\{-(t/b)^a\}.$$

Using these two distributions, independent censoring leads to a censoring rate of 14.6%.

This random censoring model was used to generate $10,000$ datasets with sample sizes $N = 20, 40$, and 60. For each of the three sample sizes, $10,000$ mid-p-values were computed using both log-rank and generalized Wilcoxon weights. Q-Q plots have been used to assess their fit to a Uniform $(0, 1)$ distribution and are shown in Figure 1. The left (right) column from top ($N = 20$) to bottom ($N = 60$) shows Q-Q plots based upon using log-rank (generalized Wilcoxon) weights.

In the important range $(0.0, 0.1)$, log-rank mid-p-values conform more closely to a uniform distribution particularly for $N = 40$ and 60 where the test based on generalized Wilcoxon weights is too liberal and rejects too often. In the $N = 20$ case, the test with log-rank weights is slightly conservative and does not reject often enough.

The accuracy of test levels subject to heavy censoring is considered in the Q-Q plots of mid-p-values in Figure 2. The tests used log-rank weights (left) and generalized Wilcoxon weights (right). Each plot used $10,000$ mid-p-values computed from datasets with sample size $N = 40$. Data were again generated from the independent censoring model but instead used a log-censoring distribution that is log-Weibull $(2, 2)$ which resulted in 40.5% censoring.

The test with log-rank weights is not rejecting often enough and is thus overly conservative with such heavy censoring. By contrast, the test with generalized Wilcoxon weights maintains accurate level in the range $(0.0, 0.1)$.

4.2 *Power*

Powers for the 5% level tests using log-rank (LGR) and generalized Wilcoxon (GWL) weights were determined when log-survival has the positively skewed extreme value distribution. This specific alternative is of interest because it suggests the use of equal log-rank weights when computing v from the log-rank class. Varying amounts of censoring and varying sample sizes were considered and the simulated powers of these tests are displayed in the entries of Table 6. For each combination of degree of

censoring $\times$ sample size, $10,000$ datasets of log-survival values were generated from the extreme value distribution and censored according to the random censoring model using a log-Weibull $(2, b)$ censoring distribution. Decreasing values of $b = 6.5, 4.0, 3.0,$ and 2.5 lead to the increasing censoring percentages from 10.1 to 42.8% that appear in the table.

The test with log-rank weights shows greater power in all settings. Both tests show deteriorating power as censoring percentage increases, as would be expected due to the loss of information. This loss in power is consistent with the views expressed by Lawless (2003, p. 38):

> "...rather large samples are often needed before the superiority of one model over another in terms of fit is indicated, and severe right censoring limits the comparison of models."

Given the more conservative level when using log-rank weights, it would appear that log-rank weights show more discriminating power against an extreme value alternative for the log-survival distribution.

5 Conclusions

The simulations suggest using log-rank weights for the test statistic rather than generalized Wilcoxon weights when testing symmetry versus positive skewness in the presence of censoring. If the data represent responses from treatment and control groups in a clinical study, and separate tests of symmetry for both treatment and control groups reject in favor of positive skewness, then greater credence should be given to the log-rank test for treatment effect rather than the generalized Wilcoxon test.

With light censoring, tests of symmetry using the log-rank weights maintained more accurate level with small and moderate sample sizes; with quite heavy censoring these tests were conservative. As concerns power, tests using log-rank weights showed greater discriminatory power between the hypotheses than tests using generalized

Wilcoxon weights. A theoretical reason for this is that log-rank weights are the consequence of the score test motivated from the positively skewed extreme value distribution and therefore lead to a test that is more responsive to an alternative hypothesis of positive skewness and the extreme value distribution in particular.

The presence of right censoring can dramatic effect our ability to detect symmetry in a distribution. With no censoring, the Gehan test is exactly the Wilcoxon test that was found to be quite powerful in the simulation study of [20]. However, when censoring is heavy, there can be a substantial loss in power due to an effective reduction in sample size as well as the loss of informativeness of censored values in the left tail from folding. A sample of n may be roughly divided in half to give $n/2$ to the left and right tail groups. With 50% censoring, roughly half the data in the left tail are uninformative about the left tail and so the effective sample sizes for the left and right tails become roughly $n/4$ and $n/2$. This explains the substantial loss in power since, with an effective sample size of $n/4$, little information remains about the left tail and this dramatically reduces the power of the test.

When approximating exact permutation significance, it should also be clear that saddlepoint mid-p-values are superior to the usual normal approximation p-values provided in the standard packages, particularly with the diminished informativeness due to splitting the tails and with censoring. Executable files with instructions for performing all five of the tests for symmetry considered above are available and may be found at

http://www.smu.edu/statistics/faculty/butler.html.

References

[1] E.F. Abd-Elfattah and R.W. Butler, *The weighted log-rank class of permutation tests: p-values and confidence intervals using saddlepoint methods*, Biometrika 94 (2007), pp. 543-551.

[2] C.C. Butler, *A test for symmetry using the sample distribution function,* Ann. Math. Statist. 40 (1969), pp. 2211-2214.

[3] G. E. Dinse, *Nonparametric estimation for partially complete time and type of failure data.* Biometrics 38 (1982), pp. 417-431.

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

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

[6] I. J. Good and R. A. Gaskins, *Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data.* J. Amer. Statist. Assoc. 75 (1980) 42-56.

[7] M.K. Gupta, *An asymptotically nonparametric test of symmetry.* Ann. Math. Statist. 38 (1967), pp. 849-866.

[8] D.L. Hill and R.V. Rao, *Tests of symmetry based on Cramér-von Mises statistics.* Biometrika 64 (1977), pp. 489-494.

[9] M. Hollander and D.A. Wolfe, *Nonparametric Statistical Methods,* Wiley, New York, 1973.

[10] J.D. Kalbfleisch and R.L. Prentice, *The Statistical Analysis of Failure Time Data*, Wiley, New York, 2002.

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

[12] J.L. Lawless, *Statistical Models and Methods for Lifetime Data,* Wiley, New York, 2003.

[13] T.P. McWilliams, *A distribution-free test for symmetry based on a runs test*, J. Amer. Statist. Assoc. 85 (1990), pp. 1130-1133.

[14] R. Peto and J. Peto, *Asymptotically efficient rank invariant test procedures*, J. Roy. Statist. Soc. A 135 (1972), pp. 185-206.

[15] S.J. Pocock, *Clinical Trials: A Practical Approach,* Wiley, New York, 1986.

[16] K.H. Pollock, S.R. Winterstein, and M.J. Cornoy, *Estimation and analysis of survival distributions for radio-tagged animals,* Biometrics 45 (1989), pp. 99-109.

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

[18] E.D. Rothman and M. Woodroofe, *A Cramér-von Mises type statistic for testing symmetry,* Ann. Math. Statist. 43 (1972), pp. 2035-2038.

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

[20] I. H. Tajuddin, *Distribution-free test for symmetry based on Wilcoxon two-sample test*, J. Appl. Statist., 21 (1994), pp. 409-416.

Table 1. Illustrative example to show centering and folding. The conversion of 2^+ to 0^+ under folding explains why censored control data do not enter into the computation of log-rank test statistics. $^+$Indicates right censoring.

Raw Data	y_i^* or (y_i^*, ∞)	Folded	(f_i, z_i, δ_i)	$f_{(i)}$	n_i	$\sum_{l \in R\{f_{(i)}\}} z_l$
1	-3	3	$(3, 0, 1)$	$f_{(3)} = 3$	1	1
$2^+ = (2, \infty)$	$-2^+ = (-2, \infty)$	$0^+ = [0, \infty)$	$(0, 0, 0)$			
3	-1	1	$(1, 0, 1)$	$f_{(1)} = 1$	4	2
$4 = \hat{S}^{-1}(0.5)$	0	0				
$5^+ = (5, \infty)$	$1^+ = (1, \infty)$	$1^+ = (1, \infty)$	$(1, 1, 0)$			
6	2	2	$(2, 1, 1)$	$f_{(2)} = 2$	2	1

Table 2. Uncentered log-times from [15]. $^+$Indicates right censoring.

0.693	1.79	2.48	3.99	4.03^+	4.22	4.49	4.56	4.56	4.83^+	4.85^+
4.88^+	4.94^+	4.95^+	4.96	4.97^+	4.98	5.00^+	5.09^+	5.12	5.15^+	5.20^+

Table 3. Mid-p-values for testing symmetry of the censored data from [15].

Mid-p-value	LGR	GWL	GH	TW	FH
Exact	.4226	.3793	.3887	.3899	.3694
Sadpt.	.4081	.3719	.3887	.3874	.3714
Normal	.2857	.3114	.3589	.3216	.3075

Table 4. Uncentered log-times from [3]. $^{+}$Indicates right censoring.

3.912	4.060	4.564	4.934	5.023	5.068	5.242	5.416	5.476	5.489
5.549	5.568	5.638	5.677	5.684	5.704^{+}	5.707	5.722^{+}	5.796^{+}	5.835^{+}
5.846^{+}	5.855^{+}	5.869^{+}	5.883	5.886^{+}	5.892^{+}	5.900^{+}	5.935^{+}	5.943^{+}	5.961^{+}

Table 5. Mid-p-values for testing symmetry of the censored data from [3].

Mid-p-value	LGR	GWL	GH	TW	FH
Exact	0.1829	0.1347	0.1185	0.1307	0.1370
Sadpt.	0.1803	0.1348	0.1183	0.1305	0.1369
Normal	0.1107	0.1124	0.1095	0.1082	0.1126

Table 6. Power of tests using log-rank (LGR) weights and generalized Wilcoxon (GWL) weights for various sample sizes and degrees of censoring. In all cases the alternative distribution for log-survival is the positively skewed extreme value distribution.

Degree of	$N = 40$		$N = 80$		$N = 160$	
Censoring	LGR	GWL	LGR	GWL	LGR	GWL
00.0%	.503	.318	.753	.456	.931	.603
10.1%	.404	.270	.572	.358	.770	.504
21.3%	.271	.202	.368	.258	.508	.352
32.8%	.192	.153	.233	.182	.304	.236
42.8%	.152	.131	.160	.141	.183	.161

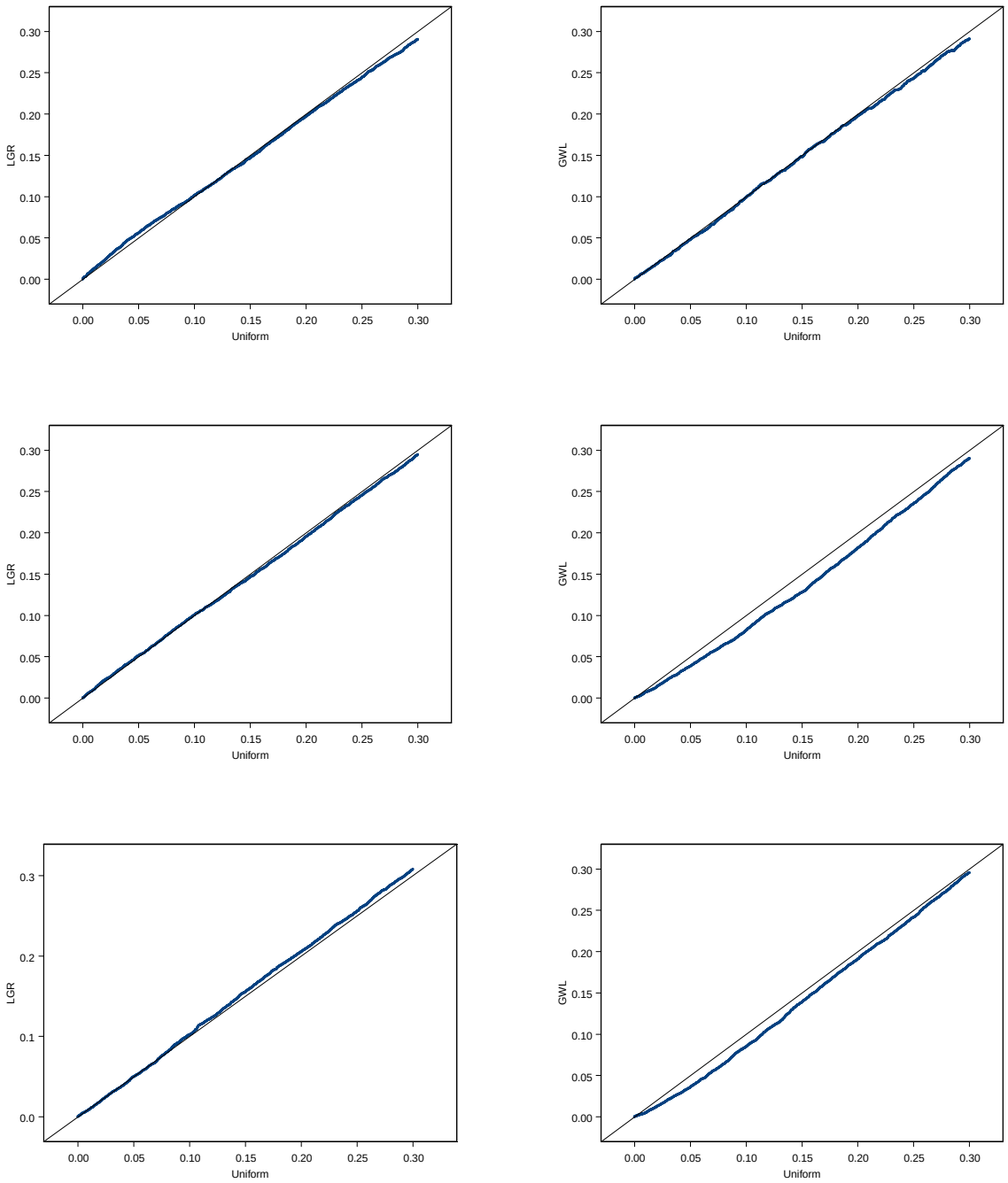

Figure 1. Q-Q plots of mid-p-values checking fit to a Uniform $(0, 1)$ distribution with light censoring (14.6%). Tests using log-rank (LGR) weights and generalized Wilcoxon (GWL) weights are shown on the left and right respectively. Rows $1, 2$ and 3 represent the increasing sample sizes $N = 20, 40$, and 60 used in the mid-p-value computation.

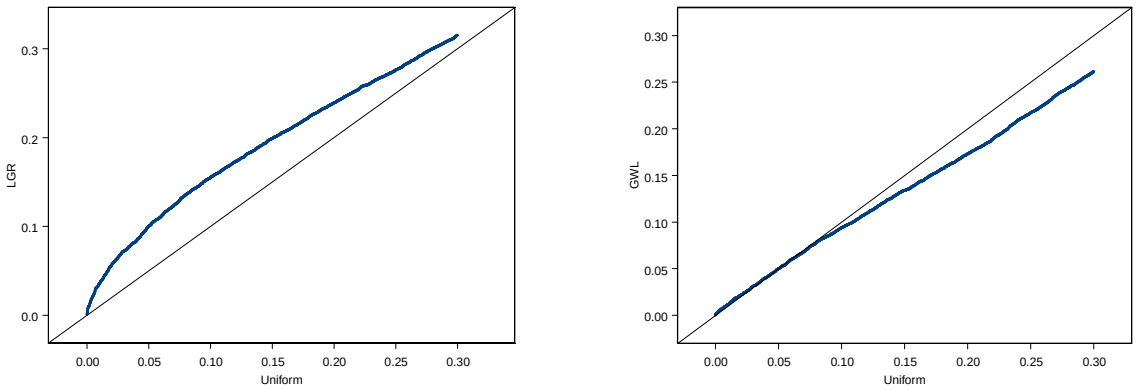

Figure 2. Q-Q plots of mid-p-values checking fit to a Uniform $(0, 1)$ distribution with heavy censoring (40.5%). Tests using log-rank (LGR) weights and generalized Wilcoxon (GWL) weights are shown on the left and right respectively. Sample size $N = 40$ was used in the mid-p-value computation.