DISSERTATION

SADDLEPOINT APPROXIMATIONS FOR LINEAR RANK MODELS

Submitted by

Ehab F. Abd-Elfattah

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2005

COLORADO STATE UNIVERSITY

May 17, 2005

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY EHAB F. ABD-ELFATTAH ENTITLED SADDLEPOINT APPROXIMATIONS FOR LINEAR RANK MODELS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN STATISTICS.

Committee on Graduate Work

(Please print name under signature)

Jose D. Salas — Jose D. Salas

M. M. Siddiqui — M. M. Siddiqui

P. W. Mielke — P. W. Mielke

R. W. Butler — R. W. Butler

Adviser

Richard A. Davis

Richard A. Davis
Department Head/Director

ABSTRACT OF DISSERTATION
SADDLEPOINT APPROXIMATIONS FOR LINEAR RANK MODELS

Linear rank tests are often used to test the effectiveness of a treatment as compares to a control in the two independent samples context. In the survival comparisons that occur in clinical trials, the time to event responses are typically right censored and modified rank tests are needed to accommodate the censoring.

This thesis proposes the use of saddlepoint approximations as a means for determining the mid-p-values for linear rank tests that involve right censoring. Normal approximations are often used by programs such as SAS to approximate the mid-p-values for these permutation distributions. These approximations are shown to be much less accurate in the simulations of this thesis. This thesis handles the weighted log-rank class of two sample tests.

The replacement of an analytical saddlepoint computation for the simulation effort required to determine percentiles of the permutation distribution makes the computation of mid-p-values simple and efficient. This speed of computation also allows for the inversion of the ranks tests to determine 95% confidence intervals from the tests. The executable files that accompany this thesis now makes the inversion of such tests routine. They deliver confidence intervals whose nominal level is, for all practical purposes, the exact intended level.

A second objective of the thesis is to deal with the test for trend when there are $k \geq 3$ treatment levels whose dosage level provides an ordering of the treatment groups. Saddlepoint approximations for the mid-p-values for the permutation distributions of the

weighted log-rank type trend tests for testing ordered alternative shown to be more accurate compare to the normal approximations.

Saddlepoint approximations also provide fast and extremely accurate methods to approximate the mid-p-values for linear rank tests for independence and symmetry. Such methods deal with the commonly used Spearman, and weighted Mann tests for independence, and Wilcoxon one sample and Wilcoxon matched pair signed rank tests for symmetry.

Ehab F. Abd-Elfattah
Department of Statistics
Colorado State University
Fort Collins, CO 80523
Summer 2005

Acknowledgments

Thanks to all who helped me to do this work. Thanks to my parents, my wife, my brothers and sisters, and my children. A special thanks to my advisor Prof. Ronald Butler for his help and comments.

I present this work to my Islamic nation as a small part of my duty to this great nation. Also as a grateful to all Egyptians who gave me from their effort and money to let me study overseas and get this degree.

Contents

Chapter 1
Introduction to the Saddlepoint Approximation

The saddlepoint method is a technique for asymptotic evaluation of integrals of the form

$$I_n = \int h(t)e^{nK(t)}dt$$

in the limit of large n, De Bruijn (1980). Integrals of this form frequently appear in probability theory, the most common source of such integrals being the inversion formula for the distribution of a sum of identically distributed random variables, in which case $h(t) = (2\pi)^{-\frac{1}{2}}e^{tx}$ and $K(t) = \ln M(t)$, where $M(t)$ is the moment generating function. For this case, the saddlepoint approximation differs from the central limit result in that, when the result is normalized so that the integrated density is equal to one, the correction term is $O(n^{-3/2})$, in contrast to the central limit theorem's $O(n^{-1/2})$, Daniels (1956). Another advantage of the saddlepoint approximation suggested by several examples is that it often leads to a uniformly bounded error, in contrast to the normal approximation in which errors generally increase in the tail of the distribution.

The saddlepoint approximation was first used in the context of probability by Cramer (1938), but its general power and scope were first suggested in an elegant paper by Daniels (1958), which contains in outline most of the major advantages of the method. Since Daniels original paper many applications have appeared in the liter-

ature. On the theoretical side, Good (1957) and (1961) discussed multidimensional generalizations and Barndorff-Nielsen and Cox (1979) presented in some detail the theory of saddlepoint expansions. Lugannani and Rice (1980) developed a systematic approach to the calculation of correction terms in the univariate distribution case. Skovgaard (1987) presented a double saddlepoint approximation for the conditional distributions in the univariate case.

Saddlepoint approximation to randomization distributions were introduced by Daniels (1954, 1958) and further developed by Robinson (1982) and Davison and Hinkley (1988). Booth and Butler (1990) showed that various randomization and resampling distributions are the same as certain conditional distributions and showed that the double saddlepoint approximation attains accuracy comparable to the single saddlepoint approach.

This chapter is an introduction to the theory of saddlepoint approximation and the formulas used throughout the following chapters.

1.1 Saddlepoint density function

Suppose a continuous random variable X has density $f(x)$ defined for all real values of x. The cumulant generating function (CGF) of f is defined as

$$K(s) = \ln M(s), \qquad s \in (a, b)$$

where $M(s)$ is the moment generating function and (a, b) is the convergent region for $M(s)$ as a neighborhood of zero. The saddlepoint density approximation to $f(x)$ is given as

$$\hat{f}(x) = \{2\pi K''(\hat{s})\}^{-1/2} \exp\{K(\hat{s}) - \hat{s}x\} \tag{1.1}$$

and $\hat{s} = \hat{s}(x)$ is the unique solution to the equation

$$K'(\hat{s}) = x \tag{1.2}$$

over the range $s \in (a, b)$. Expression (1.2) is referred to as the saddlepoint equation and $\hat{s}$ is the saddlepoint associated with value x. The approximation is meaningful for values of x that are interior points of $\chi = \{x : f(x) > 0\}$, the support of $f(x)$. We adopt the convention of referring to $\hat{f}$ as a density even though it is not really a density since $c = \int_\chi \hat{f}(x)dx \neq 1$. The normalized saddlepoint density

$$\bar{f}(x) = c^{-1}\hat{f}(x).$$

Consider, for example, $X \sim$ Gamma$(\alpha, 1)$,with density

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \qquad x > 0$$

and CGF

$$K(s) = -\alpha \ln(1-s), \qquad s \in (-\infty, 1).$$

This leads to the explicit saddlepoint $\hat{s} = 1 - \alpha/x$ for $x > 0$. So that,

$$\hat{f}(x) = (\sqrt{2\pi}\alpha^{\alpha-\frac{1}{2}}e^{-\alpha})^{-1}x^{\alpha-1}e^{-x}$$

this differs from the exact density by a replacement of the gamma function with its Stirling approximation. Thus, the normalized saddlepoint density is exact.

The derivation of the saddlepoint density approximation is often stated as it applies to the density of $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ where $X_1, ..., X_n$ are iid random variables with common CGF K. In the statement of approximation, the saddlepoint density is the leading term of an asymptotic expansion as $n \to \infty$ of the form

$$f(\bar{X}) = \hat{f}(\bar{X})\{1 + O(n^{-1})\} \tag{1.3}$$

The saddlepoint density in (1.3) is

$$\hat{f}(\bar{X}) = (2\pi K''(\hat{s})/n)^{-\frac{1}{2}} \exp\{nK(\hat{s}) - n\hat{s}\bar{X}\} \tag{1.4}$$

where $\hat{s}$ solves $K'(\hat{s}) = \bar{X}$. This expression is easily derived from (1.1), when applied to random variable $\bar{X}$.

The CGF for $n\bar{X} = \sum_{i=1}^{n} X_i$ is $nK(s)$ and defined as

$$e^{nK(s)} = \int_{-\infty}^{\infty} e^{sn\bar{X} + \ln f(\bar{X})} d\bar{X} = \int_{-\infty}^{\infty} e^{-g(s,\bar{X})} d\bar{X}$$

for $g(s,\bar{X}) = -sn\bar{X} - \ln f(\bar{X})$. With s fixed, Laplace's approximation for the integral is

$$e^{nK(s)} \simeq \sqrt{\frac{2\pi}{g''(s,\bar{X}_s)}} e^{sn\bar{X}_s} f(\bar{X}_s) \tag{1.5}$$

where $g''(s,\bar{X}) = \partial^2 g(s,\bar{X})/\partial\bar{X}^2 = -\partial^2 \ln f(\bar{X})/\partial\bar{X}^2$ and $\bar{X}_s$minimizes $g(s,\bar{X})$ over $\bar{X}$. As a critical value $\bar{X}_s$ solves

$$0 = -g'(s,\bar{X}_s) = ns + \frac{\partial \ln f(\bar{X}_s)}{\partial \bar{X}_s}$$

partial differentiation with respect to $\bar{X}_s$ gives

$$\frac{\partial s}{\partial \bar{X}_s} = -\frac{1}{n}\frac{\partial^2 \ln f(\bar{X}_s)}{\partial \bar{X}_s^2}$$

i.e. $g''(s,\bar{X}_s) = -\partial^2 \ln f(\bar{X}_s)/\partial\bar{X}_s^2 = n\partial s/\partial\bar{X}_s = n(\partial\bar{X}_s/\partial s)^{-1} = n\{K''(s)\}^{-1}$. The last term is determined by differentiation of the saddlepoint equation. The expression (1.5) becomes

$$e^{nK(s)} \simeq \sqrt{\frac{2\pi K''(s)}{n}} e^{sn\bar{X}_s} f(\bar{X}_s)$$

or

$$f(\bar{X}_s) \simeq \sqrt{\frac{n}{2\pi K''(s)}} \exp\{nK(\hat{s}) - n\hat{s}\bar{X}\}$$

1.2 Cumulative Distribution Function

Suppose continuous random variable X has CDF $F(x)$ and CGF $K(s)$ with mean $\mu = E(X)$. The saddlepoint approximation for $F(x)$, as introduced in Lugannani and Rice (1980), is

$$\widehat{F}(x) := \begin{cases} \Phi(\hat{w}) + \phi(\hat{w})(1/\hat{w} - 1/\hat{u}) & if\ x \neq \mu \\ \frac{1}{2} + K'''(0)/\{6\sqrt{2\pi}K''(0)^{3/2}\} & if\ x = \mu \end{cases} \tag{1.6}$$

where

$$\begin{aligned} \hat{w} &= \text{sgn}(\hat{s})\sqrt{2\{\hat{s}x - K(\hat{s})\}} \\ \widehat{u} &= \hat{s}\sqrt{K''(\hat{s})} \end{aligned} \tag{1.7}$$

and $\hat{s}$ is the unique solution to $K'(\hat{s}) = x$. Symbols ϕ and Φ denote the standard normal density and CDF respectively. The continuous CDF approximation is derived below based on an approximation due to Temme (1982). The approach involves first expressing the CDF approximation as the finite cumulative integral of the saddlepoint density, then, using the Temme approximation to approximate it.

Definition *Suppose Z_{w_0} has a normal (0,1) distribution truncated to attain values no larger than w_0. The Temme approximation gives an approximation for the numerator in the computation of $E\{h(Z_{w_0})\}$ of the form*

$$\int_{-\infty}^{w_0} h(w)\phi(w)dw \simeq h(0)\Phi(w_0) + \phi(w_0)\{\frac{h(0) - h(w_0)}{w_0}\}$$

The derivation of (1.6) proceeding by integrating the saddlepoint density

$$F(y) \simeq \int_{-\infty}^{y} \hat{f}(X)dX = \int_{-\infty}^{y} \{2\pi K''(\hat{s})\}^{-1/2} \exp\{K(\hat{s}) - \hat{s}X\}dX \quad (1.8)$$
$$= \int_{-\infty}^{y} K''(\hat{s})^{-\frac{1}{2}} \phi(\hat{w})dX$$

where $\hat{w}$ is defined as in (1.7) and $\hat{w}$ and $\hat{s}$ are functions of X. A change of variables in (1.8) from dX to $d\hat{w}$ puts the integral in a form for which the Temme approximation can be applied. The differential of the mapping $X \leftrightarrow \hat{w}$ is computed by the following lemma.

Lemma *The mapping $X \leftrightarrow \hat{w}$ is smooth monotonic increasing transformation with derivative*

$$\frac{dX}{d\hat{w}} = \begin{cases} \hat{w}/\hat{s} & if\ \hat{s} \neq 0 \\ K''(0)^{1/2} & if\ \hat{s} = 0 \end{cases}$$

Proof. See Butler (2005).

Implementing this change in (1.8) yields

$$\int_{-\infty}^{y} K''(\hat{s})^{-\frac{1}{2}} \phi(\hat{w})dX = \int_{-\infty}^{\hat{w}_y} \frac{\hat{w}}{\hat{s}\sqrt{K''(\hat{s})}} \phi(\hat{w})d\hat{w}$$

where $\hat{w}_y$ is the value of $\hat{w}$ determined by solving the saddlepoint equation at y. The Temme approximation is applied, with $h(\hat{w}) = \hat{w}/\{\hat{s}\sqrt{K''(\hat{s})}\}$. The value $h(0)$ is determined by using the lemma

$$h(0) = \lim_{\hat{s}\to 0} \hat{w}/\{\hat{s}\sqrt{K''(\hat{s})}\} = 1$$

So (1.8) becomes

$$\hat{F}(y) = \Phi(\hat{w}_y) + \phi(\hat{w}_y)(\frac{1 - h(\hat{w}_y)}{\hat{w}_y}).$$

Discrete CDF approximation requires modification to the formula for the continuous CDF approximation in order to achieve the greatest accuracy. Daniels (1987) introduced two such continuity corrected modifications.

The saddlepoint density in the multivariate case for continuous m-dimension vector X is defined on the interior of the convex hull of the support (I_χ) of f as

$$\hat{f}(X) = (2\pi)^{-m/2} K''(\hat{s})^{-1/2} \exp\{K(\hat{s}) - \hat{s}^T X\}, \qquad X \in I_\chi \tag{1.9}$$

where $\hat{s}$ is the unique solution to the m-dimensional saddlepoint equation $K'(\hat{s}) = X$.

1.3 Conditional Saddlepoint Density Function

Let (X, Y) be a random vector having a non-degenerate distribution in $\Re^m$ with $\dim(X) = m_x$, $\dim(Y) = m_y$ and $m_x + m_y = m$. Suppose there is a joint density $f(x, y)$ with support $(x, y) \in \chi \subseteq \Re^m$. The conditional density of Y at y given $X = x$ is defined as

$$f(y|x) = \frac{f(x, y)}{f(x)}, \qquad (x, y) \in \chi$$

and 0 otherwise. A natural approximation for $f(y|x)$ is to use two separate saddlepoint approximations for $f(x, y)$ and $f(x)$. Denoting such an approximation with the inclusion of hats, then

$$\hat{f}(y|x) = \frac{\hat{f}(x, y)}{\hat{f}(x)}, \qquad (x, y) \in I_\chi \tag{1.10}$$

defines a double saddlepoint density. The idea of using two saddlepoint approximations to recover a conditional density was introduced by Daniels (1958) with its full elaboration presented in Barndorff-Nielsen and Cox (1979).

Double saddlepoint density (1.10) may be expressed in terms of the joint CGF, $K(x, y)$ as follows. First write

$$\hat{f}(x,y) = (2\pi)^{-m/2} \left|K''(\hat{s},\hat{t})\right|^{-1/2} \exp\{K(\hat{s},\hat{t}) - \hat{s}^T x - \hat{t}^T y\} \tag{1.11}$$

where the m-dimensional saddlepoint $(\hat{s}, \hat{t}\,)$ solves the set of m equations

$$K'(\hat{s},\hat{t}) = (x,y) \tag{1.12}$$

with K' as the gradient with respect to both components s and t, and K'' as the corresponding Hessian. The denominator saddlepoint density is determined from the marginal CGF of X, given as $K(s, 0)$, and is

$$\hat{f}(x) = (2\pi)^{-m_x/2} \left|K''_{ss}(\hat{s}_0,0)\right|^{-1/2} \exp\{K(\hat{s}_0,0) - \hat{s}_0^T x\} \tag{1.13}$$

where $\hat{s}_0$ is the m-dimensional saddlepoint for the denominator that solves $K'_s(\hat{s}_0, 0) = x$. Putting (1.11) and (1.12) together gives a convenient computational form as

$$\hat{f}(y|x) = (2\pi)^{-m_y/2} \left\{ \frac{\left|K''(\hat{s},\hat{t})\right|}{\left|K''_{ss}(\hat{s}_0,0)\right|} \right\}^{-1/2} \exp[\{K(\hat{s},\hat{t}) - \hat{s}^T x - \hat{t}^T y\} - \{K(\hat{s}_0,0) - \hat{s}_0^T x\}] \tag{1.14}$$

1.4 Conditional Cumulative Distribution Function

A double saddlepoint CDF approximation for Y given $X = x$ when $\dim(Y) = 1$ and $\dim(X) = m$ has been given in Skovgaard (1987). Suppose that

$$F(y|x) = \Pr(Y \leq y|X = x)$$

is the true CDF. The Skovgaard approximation is

$$\hat{F}(y|x) := \Phi(\hat{w}) + \phi(\hat{w})(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}), \qquad \hat{t} \neq 0 \tag{1.15}$$

where

$$\begin{aligned} \hat{w} &= \text{sgn}(\hat{t})\sqrt{2[\{K(\hat{s}_0, 0) - \hat{s}_0^T x\} - \{K(\hat{s}, \hat{t}) - \hat{s}^T x - \hat{t} y\}]} \\ \hat{u} &= \hat{t}\sqrt{\frac{|K''(\hat{s}, \hat{t})|}{|K''_{ss}(\hat{s}_0, 0)|}} \end{aligned}$$

The proof follows the same approach used in deriving the Lugannani and Rice approximation. The idea is to approximate the left tail probability as the integral of the double saddlepoint density using the Temme procedure.

An approximation to $F(z|x)$ is the integral of the double saddlepoint density as

$$\int_{-\infty}^{z} \hat{f}(y|x)dy = \int_{-\infty}^{z} \left\{ \frac{|K''(\hat{s}, \hat{t})|}{|K''_{ss}(\hat{s}_0, 0)|} \right\}^{-1/2} \phi(\hat{w})dy$$

where $\hat{w}$ is defined in (1.7). First the change of variables $y \leftrightarrow \hat{w}$ is required with Jacobian in the given in the lemma.

Lemma *The mapping $y \leftrightarrow \hat{w}$ is smooth and monotonic increasing with derivative*

$$\frac{dy}{d\hat{w}} = \begin{cases} \hat{w}/\hat{t} & if \ \hat{t} \neq 0 \\ \sqrt{|K''(\hat{s}_0, 0)| \,/\, |K''_{ss}(\hat{s}_0, 0)|} & if \ \hat{t} = 0 \end{cases}$$

Proof. See Butler (2005).

Using this Lemma,

$$\int_{-\infty}^{z} \hat{f}(y|x)dy = \int_{-\infty}^{z} \left\{ \frac{|K''(\hat{s}, \hat{t})|}{|K''_{ss}(\hat{s}_0, 0)|} \right\}^{-1/2} \frac{\hat{w}}{\hat{t}} \phi(\hat{w}) d\hat{w}$$

and by the Temme approximation

$$\int_{-\infty}^{z} \hat{f}(y|x)dy = \Phi(\hat{w}_z) + \phi(\hat{w}_z)(\frac{1 - h(\hat{w}_z)}{\hat{w}_z})$$

where $h(\hat{w}_z) = \left\{ |K''(\hat{s}, \hat{t})| \,/\, |K''_{ss}(\hat{s}_0, 0)| \right\}^{-1/2} \hat{w}/\hat{t}$ and $h(0) = 1$

The asymptotic accuracy of the approximation, as the leading term of an expansion, was given by Skovgaard and is similar to that of the Lugannani and Rice approximation. If $(\bar{x}, \bar{y})$ remains in a large deviation set as $n \to \infty$, then

$$F(\bar{y}|\bar{x}) = \hat{F}(\bar{y}|\bar{x}) + O(n^{-1})$$

in both the continuous and lattice cases. Over sets of bounded central tendency, the $O(n^{-1})$ error improves to $O(n^{-3/2})$.

Chapter 2
Saddlepoint Tests and Confidence Intervals for the Log-Rank and Generalized Wilcoxon Rank Tests

2.1 Introduction

Linear rank tests are often used to test the effectiveness of a treatment as compares to a control in the two independent samples context. In the survival comparisons that occur in clinical trials, the time to event responses are typically right censored and modified rank tests are needed to accommodate the censoring. The most commonly used modified rank tests are the log-rank and generalized Wilcoxon tests (Gehan, 1965a and 1965b; Breslow, 1970; Peto and Peto, 1972; Prentice, 1978). These tests are most conveniently motivated as arising from the accelerated failure time model as described in Kalbfleish and Prentice (2002).

The current paper proposes the use of saddlepoint approximations as a means for determining the significance levels for these two tests under their exact permutation distributions. Currently, programs such as SAS use asymptotic normal approximations as described, for example, in Prentice (1978) and Kalbfleish and Prentice (2002). The saddlepoint approximations, however, are almost always closer to the true permutation significance levels and the degree of their greater accuracy is read-

ily apparent in smaller and intermediate size samples for which asymptotic normality has not been attained. A variety of examples and simulations are used to show that the saddlepoint mid-p-value is an extremely accurate approximation for the mid-p-value as determined from the exact permutation distribution.

The focus is on mid-p-values rather than ordinary p-values because a major aim of this paper is to construct confidence intervals for the treatment effect through the inversion of the tests. When mid-p-values are used instead of ordinary p-values in this inversion, the intervals that result have true converges that tend to be much closer to the nominal coverage. This fact been discussed extensively in Agresti (1992), Kim and Agresti (1995), and Butler (2005, chapter 6). When the mid-p-values determined by saddlepoint approximation are inverted, they lead to confidence intervals that are almost identical to the exact intervals determined by difficult simulations of the exact permutation distributions. Confidence intervals for the inversion of normal tests are described in Kalbfleish and Prentice (2002) and are consistently less accurate. A extensive range of practical examples is considered and in each case the saddlepoint intervals almost exactly replicate the true intervals from the exact permutation distribution.

The treatment of ties requires a separate discussion in §4. The two standard remedies use score average methods or permutation methods. In either case, the determination of mid-p-values by using saddlepoint approximations displays unrivalled accuracy.

2.2 Linear Rank Tests

An independent sample of n_1 responses is observed for a treatment group independently of another independent sample of n_2 responses from a control group. Each sample is subject to independent censoring according to the standard assumptions that state that the chance of censoring and the censoring distribution are not dependent on group membership. Denoting the uncensored survival functions for treatment and control as $S_1(t)$ and $S_2(t)$ respectively, then a test for treatment effectiveness, or $H_0 : S_1(t) \equiv S_2(t) = S(t)$ versus $H_1 : S_1(t) \neq S_2(t)$ for some t is generally based on rank tests such as the log-rank and the generalized Wilcoxon rank tests.

2.2.1 Accelerated Failure Time Model

The accelerated failure time (AFT) model is most often used to derive the scores of these rank tests, although it was not the original context in which these tests were developed. Consider first the setting with no censoring and no ties in the survival times. Suppose that $n = n_1+n_2$ and the data for patient i are (z_i, t_i) where $z_i \in \{0, 1\}$ is an indicator of treatment group membership, and t_i is the survival time. The AFT model with a single independent variable $z = (z_1, ..., z_n)^T$ is the linear model

$$y = 1\alpha + z\beta + \sigma e \tag{2.1}$$

where $y = (y_1, ..., y_n)^T = (\log t_1, ..., \log t_n)^T$, $1 = (1, ..., 1)^T$ is $n \times 1$, and $e = (e_1, ..., e_n)^T$ consists of i.i.d. errors from an unknown distribution F. If r_i is the rank of y_i and $r = (r_1, ..., r_n)^T$, then a nonparametric test can be based on ranks r. The

locally most powerful rank test for testing $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ uses $\partial \log p(r)/\partial\beta|_{\beta=0}$ where $p(r)$ is the probability of rank vector r that depends on F.

With the additional complication of independent censoring, Peto and Peto (1972) and Prentice (1978) have determined locally most powerful rank tests for the same hypothesis where the particular test depends on F. In this context the data for patient i are (z_i, t_i, δ_i) with the additional variable δ_i to indicate whether patient i has been censored. With $\delta = (\delta_1, ..., \delta_n)^T$ and r again the the vector of ranks, they derive the locally most powerful rank test based on $\partial \log p(r, \delta)/\partial\beta|_{\beta=0}$ where $p(r, \delta)$ is the probability for the rank and censoring vectors. The test has the form $v = \sum_{i=1}^{n} w_i z_i$ where weight w_i depends on δ_i and characteristics of the distribution F.

Log-Rank and Generalized Wilcoxon Tests

The log-rank or modified Savage exponential scores test results from assuming that F is an extreme value distribution or Gumbel. Ties in survivals are again not allowed although the means for dealing with them are considered later in §4. A description of the test is made considerably simpler by changing the notation to that used in Kalbfleish and Prentice (2002). Let $t_1, ..., t_k$ denote the observed survival times for both groups so that $\{t_i\}$ no longer include the censoring times. Suppose these survival times have the corresponding covariates $z_1, ..., z_k$ and let $y_i = \log t_i$. Order the log-survival times as $y_{(1)} < \cdots < y_{(k)}$ with the corresponding covariates $z_{(1)}, ..., z_{(k)}$. To deal with the censored values interspersed among the order statis-

tics, suppose that $y_{i1}, ..., y_{im_i}$ are censored values in $[y_{(i)}, y_{(i+1)})$, $i = 0, ..., k$ whose corresponding covariates are $z_{i1}, ..., z_{im_i}$. Linear rank statistics take the form

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

where c_i and C_i are the scores for the uncensored and censored data points and their values determine the specific test.

When F has an extreme value distribution, Prentice (1978) has shown that the log-rank test, as originally suggested in Peto an Peto (1972), is the locally most powerful rank test. The score values are

$$c_i = \sum_{j=1}^{i} n_j^{-1} - 1, \qquad C_i = \sum_{j=1}^{i} n_j^{-1}, \tag{2.3}$$

where n_j indicates the number of patients at risk at time $t_{(j)}^-$.

The permutation distribution will only concern the structure of v through (2.3). For the asymptotic normal approximation, an estimate of the asymptotic variance for standardization is given by Kalbfleish and Prentice (2002) as

$$V_0 = \sum_{i=1}^{k} n_i^{-1} \sum_{l \in R\{t_{(i)}\}} (z_l - \bar{z}_i)^2, \qquad \bar{z}_i = n_i^{-1} \sum_{l \in R\{t_{(i)}\}} z_l, \tag{2.4}$$

where $R\{t_{(i)}\}$ consists of the indices for those individuals at risk at time t_i^-, the so called risk set at time t_i^-.

When the logistic distribution is used as the error distribution F in the AFT model, then Prentice (1978) has shown that the generalized Wilcoxon test is the locally most powerful test. This test has the same structure but with the alternative

scores

$$c_i = 1 - 2\prod_{j=1}^{i} \frac{n_j}{n_j + 1}, \qquad C_i = 1 - \prod_{j=1}^{i} \frac{n_j}{n_j + 1}. \tag{2.5}$$

The asymptotic normal approximation for the standardized value of v uses the variance estimate

$$V_0 = \sum_{i=1}^{k} \left[a_i(1 - a_i^*) \left\{ 2z_{(i)}^2 + \sum_{j=1}^{m_i} z_{ij}^2 \right\} - (a_i^* - a_i)x_{(i)} \left\{ a_i x_{(i)} + 2 \sum_{j=i+1}^{k} a_j x_{(j)} \right\} \right], \tag{2.6}$$

where

$$a_i = \prod_{j=1}^{i} \frac{n_j}{n_j + 1}, \qquad a_i^* = \prod_{j=1}^{i} \frac{n_j + 1}{n_j + 2}, \qquad x_{(i)} = 2z_{(i)} + s_{(i)}, \qquad s_{(i)} = \sum_{j=1}^{m_i} z_{ij}.$$

2.2.2 Saddlepoint Approximation

When considering the permutation distribution of statistic v in (2.2), the sequence of z_i variables is generally assumed to have the distribution of a random permutation vector $H = (H_1, ..., H_n)^T$ with n_1 ones and n_2 zeros. Thus any one of the distinct $\binom{n}{n_1}$ permutations for H is assumed to have probability $1/\binom{n}{n_1}$. The null distribution for v is determined by its linear dependence on the permutation vector H. The fact that this dependence is linear in H leads to the following characterization that makes it amenable to double saddlepoint approximation as shown in Booth and Butler (1990).

Theorem 1 *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 H. This provides the conditional characterization for the null distribution of v as

$$P = \sum_{i=1}^{n} q_i H_i \sim \sum_{i=1}^{n} q_i Z_i \quad \textit{given} \quad \sum_{i=1}^{n} Z_i = n_1.$$

The weights $\{q_i\}$ are weights $\{c_i\}$ for survival times and weights $\{C_i\}$ for censored times. Thus, when considered in the applications, the weights $\{q_i\}$ depend on the relative placement of the censored values in the ordered sample.

The Skovgaard (1987) saddlepoint approximation leads to a simple conditional probability computation that is based on the conditional characterization in Theorem 1. Let

$$X = \sum_{i=1}^{n} Z_i \qquad \text{and} \qquad Y = \sum_{i=1}^{n} q_i Z_i$$

and compute the joint MGF as

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

Let $K(s,t) = \log M_{X,Y}(s,t)$ be the joint cumulant generating function (CGF). If v_0 is the observed value of the statistic (2.2), the p-value from the Skovgaard approximation is

$$\Pr(P \geq v_0) = \Pr(Y \geq v_0 | X = n_1) \simeq 1 - \Phi(\hat{w}) - \phi(\hat{w}) \left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}} \right) \qquad (2.7)$$

where

$$\hat{w} = \operatorname{sgn}(\hat{t}) \sqrt{2 \left[\{K(\hat{s}_0, 0) - n_1 \hat{s}_0\} - \{K(\hat{s}, \hat{t}) - n_1 \hat{s} - v_0 \hat{t}\} \right]}$$

$$\hat{u} = \hat{t} \sqrt{\left| K''(\hat{s}, \hat{t}) \right| / K''_{ss}(\hat{s}_0, 0)}.$$

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

$$K'_s(\hat{s}, \hat{t}) = \sum_{i=1}^{n} \frac{\exp(\hat{s} + q_i\hat{t})}{(1-\theta)/\theta + \exp(\hat{s} + q_i\hat{t})} = n_1$$
$$K'_t(\hat{s}, \hat{t}) = \sum_{i=1}^{n} \frac{q_i \exp(\hat{s} + q_i\hat{t})}{(1-\theta)/\theta + \exp(\hat{s} + q_i\hat{t})} = v_0$$

while the denominator saddlepoint $\hat{s}_0$ solves

$$K'_s(\hat{s}_0, 0) = \frac{n \exp(\hat{s}_0)}{(1-\theta)/\theta + \exp(\hat{s}_0)} = n_1. \tag{2.8}$$

Since the computations of $\hat{w}$ and $\hat{u}$ do not depend on the particular value of θ used, the value $\theta = n_1/n$ has been used since it leads to the explicit solution for (2.8) as $\hat{s}_0 = 0$ and provides some simplification to the calculations.

The saddlepoint expression in (2.7) has been formally derived for continuous P but in the permutation setting P is discrete and not even a lattice distribution for which a continuity correction would be available. An explanation is necessary as to why this continuous formula can and should be used. The answer is that it provides the most accurate approximation for the mid-p-value given by

$$\text{mid-}p := \Pr(P > v_0) + \tfrac{1}{2}\Pr(P = v_0), \tag{2.9}$$

which is the p-value deflated by half the probability for $\Pr(P = v_0)$. Agresti (1992), Routledge (1994) and Kim and Agresti (1995) have advocated its use over the p-value claiming that the ordinary p-value is too conservative. This claim finds its strongest justification when the significance tests are inverted to provide confidence intervals for the parameters under test. The use of 2.5% mid-p-values in either tail

leads to confidence intervals whose nominal coverage is 95% and whose attained coverage is extremely close to this nominal coverage. This is not the case when the p-value is used to invert the test. When p-values are inverted the attained coverage tends to be consistently larger than the intended nominal coverage particularly with smaller sample sizes. Thus, the use of the p-value is inherently too conservative in the construction of confidence intervals from tests. Examples showing the superiority of using the mid-p-value are given in the references above as well as Chapman, Butler, and Paige (2005) who provide saddlepoint confidence intervals for LD-50 in logistic regression.

The determination of confidence intervals with the correct coverage in §3 is one of our main goals and so the presentation focusses on mid-p-values to construct intervals with more accurate coverage. Furthermore, if there is going to be any consistency in interpretation between significance levels and coverage probabilities, then it is the mid-p-value that needs to be used for the significance computation.

The fact that continuous saddlepoint procedures tend to approximate mid-p-values in discrete settings has been discussed in Pierce and Peters (1992), Davison and Wang (2002), and Butler (2005, §6.1.4). The simplest explanation is given in the last reference and takes the view that the continuous saddlepoint approximation is simply an approximation to the true Fourier inversion, e.g. the true survival function. The true survival function $\Pr(P \geq v_0)$ has a step discontinuity at v_0, and it is known that exact Fourier inversion yields the midpoint of the step in the discontinuity or the

mid-p-value. Thus, the interpretation of the continuous saddlepoint approximation in the permutation distribution context is that it approximates the true Fourier inversion thus providing an approximation to the mid-p-value as displayed in (2.9).

2.2.3 Illustration

Simple illustrations of the saddlepoint method can be based on two published sets of data. The first data set is smaller and is given in Kalbfleisch and Prentice (2002, pg. 222) with $n_1 = 4$ and $n_2 = 5$. The second larger set of data was used by Pike (1966) and Prentice (1978) and records the time until development of vaginal cancer for two groups of female rats of size $n_1 = 21$ and $n_2 = 19$ after exposure to a carcinogen. In both examples, the hypothesis test is that there is no group difference, or $\beta = 0$ in the AFT model treatment, versus treatment (group 1) survives longer than control or $\beta > 0$. The second data set has some ties occurring within the same group but this does not affect the two test statistics; the same test statistic values result no matter which way the ties are ordered.

For these two data sets, Table 1 summarizes the computation of normal p-values (which are naturally approximations for mid-p-values), saddlepoint mid-p-values, and the true (simulated) mid p-values for the log-rank and generalized Wilcoxon tests. The top four rows treat the censored values as actual survival times so that no censored values need to be accounted for. In each instance, the true mid-p-value has been calculated by taking 10^6 simple random samples of n_1 from n, holding the cen-

soring orders fixed, and computing the proportion of times that P exceeds v_0 plus half the proportion of time it equals v_0. With larger data sets the distinction between mid-p-value and p-value becomes negligible since the mass at v_0 is quite small.

Data	n_1	n_2	Test stat.	Normal p	Sadpt. mid-p	True[1] mid-p
All items treated as uncensored.						
Set 1	4	5	LR	0.01590	0.02159	0.01979
			GW	0.02213	0.02327	0.02388
Set 2	21	19	LR	0.03064	0.03458	0.03462
			GW	0.04637	0.04802	0.04745
With Censored Items						
Set 1	4	5	LR	0.00937	0.01725	0.01580
			GW	0.01164	0.01356	0.01581
Set 2	21	19	LR	0.04875	0.05636	0.05686
			GW	0.04960	0.05226	0.05242

Table 1. Normal, saddlepoint and true mid-p-values for the log-rank (LR) and generalized Wilcoxon (GW) statistics applied to the two sets of data. [1]Based on 10^6 simple random samples of n_1 from n and holding the censoring orders fixed.

Table 1 shows that the saddlepoint approximation is highly accurate for both the smaller and larger data sets and also with and without censoring. By contrast, the normal approximation only works well with no censoring and shows inaccuracy with censoring even for the larger data set 2.

2.2.4 Simulation Study

The saddlepoint accuracy seen in Table 1 occurs consistently over a wide range of conditions. Simulation studies are used to demonstrate this consistent performance for a range of sample sizes, degrees of sample imbalance, and prevalence of censoring. The saddlepoint methods are extremely accurate for all the settings considered.

Three error distributions are used to simulate data: Logistic, extreme value and Weibull distributions. For each distribution, 1000 data sets were generated in the following way. First $n = n_1 + n_2$ i.i.d. responses were drawn from the distribution. Second, n_1 of these values were selected at random to determine the locations for treatment. Third, the treatment values were translated to the right an amount $\beta > 0$ designed to induce borderline significance of a treatment effect, since this really is the most interesting situation. Finally a preset number of observations from each group were relabelled as censored. Small, intermediate and large sample sizes were used which were either balanced or unbalanced among the groups. The censoring percentage also changed between light, intermediate and heavy censoring. The aim in these choices was to keep the mean mid-p-value near 0.05.

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

Stat.	Mean	Sad. Prop.	Abs. Err. Sad.	Abs. Err. Normal	Rel. Abs. Err. Sad	Rel. Abs. Err. Nor.
$n_1 = 18$		$n_2 = 17$	$\beta = 1.5$		30% censoring	
LR	.145	.983	$.0^3484$	.0128	$.0^4100$	$.0^3167$
GW	.108	.980	$.0^3198$	$.0^2436$	$.0^5300$	$.0^3114$
$n_1 = 18$		$n_2 = 17$	$\beta = 1.5$		5% censoring	
LR	.059	.996	$.0^3413$	$.0^2353$	$.0^5800$	$.0^3115$
GW	.041	.855	$.0^3132$	$.0^3810$	$.0^5100$	$.0^4160$
$n_1 = 8$		$n_2 = 7$	$\beta = 2.0$		15% censoring	
LR	.117	0.995	$.0^2137$	.0237	$.0^4140$	$.0^3459$
GW	.090	0.982	$.0^3403$	$.0^2663$	$.0^5830$	$.0^4630$
$n_1 = 36$		$n_2 = 34$	$\beta = 1.0$		5% censoring	
LR	.122	.986	$.0^3343$	$.0^2509$	$.0^6100$	$.0^4130$
GW	.087	.909	$.0^3178$	$.0^2111$	$.0^5444$	$.0^4300$

Table 2. Performance under simulation from the logistic distribution.

Consider for example the simulation of 1000 data sets with $n_1 = 18$ and $n_2 = 17$ and 30% censoring. With the log-rank test, the saddlepoint mid-p-value was closer to the true value 98.3% of the time. For the generalized Wilcoxon test, locally most powerful with this logistic simulation, the saddlepoint approximation was closer 98% of the time. For log-rank test, the absolute error of the normal approximation was 1.28% versus 0.0484% for the saddlepoint approximation, and the relative error for the normal was .0167% versus .001% for the saddlepoint approximation.

Stat.	Mean	Sad. Prop.	Abs. Err. Sad.	Abs. Err. Normal	Rel. Abs. Err. Sad	Rel. Abs. Err. Nor.
$n_1 = 12$		$n_2 = 8$	$\beta = 0.8$		30% censoring	
LR	.085	.999	$.0^2110$	.0339	$.0^4254$	$.0^3284$
GW	.046	.997	$.0^3296$	.0108	$.0^6128$	$.0^3580$
$n_1 = 18$		$n_2 = 17$	$\beta = 0.5$		30% censoring	
LR	.054	.987	$.0^3376$	$.0^2502$	$.0^5219$	$.0^3221$
GW	.028	.872	$.0^4886$	$.0^3980$	$.0^5794$	$.0^3145$
$n_1 = 36$		$n_2 = 34$	$\beta = 0.3$		30% censoring	
LR	.09	.987	$.0^3221$	$.0^2387$	$.0^6266$	$.0^5273$
GW	.06	.914	$.0^3150$	$.0^2110$	$.0^5317$	$.0^4377$
$n_1 = 30$		$n_2 = 10$	$\beta = 0.55$		30% censoring	
LR	.055	1.0	$.0^3303$	.0186	$.0^4220$	$.0^3472$
GW	.032	1.0	$.0^3106$	$.0^2740$	$.0^5400$	$.0^3486$

Table 3. Performance under simulation from the Weibull distribution.

Overall, the saddlepoint approximation performed better than the normal approximation in all cases and the discrepancy was greater for the log-rank than the generalized Wilcoxon test. When averaged over all simulations, the saddlepoint approximation was closer 99.18% and 91.95% of the time respectively for the log-rank and generalized Wilcoxon tests. In most cases the saddlepoint approximation demonstrated a relative error that is less than .01% and an absolute deviation of 1% . Unbalanced data, heavy censoring, and nonsymmetric error distributions had a detrimental effect on the accuracy of the normal approximation.

Stat.	Mean	Sad. Prop.	Abs. Err. Sad.	Abs. Err. Normal	Rel. Abs. Err. Sad	Rel. Abs. Err. Nor.
$n_1 = 8$	$n_2 = 7$	$\beta = 1.5$			15% censoring	
LR	.049	1.0	$.0^3895$	.0122	$.0^5100$	$.0^3389$
GW	.055	.83	$.0^3764$	$.0^2226$	$.0^5900$	$.0^4290$
$n_1 = 18$	$n_2 = 17$	$\beta = 1.0$			15% censoring	
LR	.031	1.0	$.0^3247$	$.0^2435$	$.0^4230$	$.0^3607$
GW	.044	.860	$.0^3121$	$.0^3793$	$.0^4230$	$.0^4220$
$n_1 = 32$	$n_2 = 38$	$\beta = 0.7$			30% censoring	
LR	.040	.979	$.0^3128$	$.0^2209$	$.0^4160$	$.0^3267$
GW	.055	.856	$.0^3132$	$.0^3746$	$.0^4110$	$.0^4650$
$n_1 = 10$	$n_2 = 30$	$\beta = 1.0$			30% censoring	
LR	.048	.990	$.0^3212$	$.0^2646$	0^4180	$.0^3490$
GW	.064	.979	$.0^3147$	$.0^2403$	$.0^5400$	$.0^3223$

Table 4. Performance under simulation from the extreme value distribution.

2.3 Confidence Interval for β

The tests of significance described in §2.2 can be inverted to determine 95% confidence intervals for the unknown parameter β. The standard correspondence of tests and confidence intervals is used in which the rank test of $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ has a mid-p-value within the range $[.025, .975]$ if and only if β_0 is in the 95% confidence interval.

To perform the rank test of $H_0 : \beta = \beta_0 \neq 0$ within the framework of the AFT model in (2.1), simply use the previous rank tests on the $n \times 1$ vector of residuals $y - z\beta_0$ rather than y. Prentice (1978) has inverted such tests by using the asymptotic normal test statistic. Using a fine grid of β_0-values with increment 0.01, he computed

normal p-values for the log-rank and generalized Wilcoxon tests. As a function of increasing values of β_0, the normal test statistic $v/\sqrt{V_0}$ is a step function that makes an incremental decrease whenever the residual for a treatment subject is interchanged with the value of a control subject. When a 0.01 increment in the value of β_0 does not lead to an interchange, then the normal statistic and p-value remain unchanged and the dependence is flat.

The idea of inverting rank tests is conceptually simple and easy to implement, however there are some subtleties that need to be noted. First consider the determination of the required cutoff v_0 and, for purposes of discussion, suppose that $\beta_0 > 0$. For the treatment group all survival and censored times are diminished by amount β_0 which changes the relative ordering of treatment and control responses as well as the ordering of the positions held by censored observations. Denote the determination of the observed test statistic with treatment translation β_0 as $v_0(\beta_0)$. As previously mentioned with the normal approximation, the cutoff $v_0(\cdot)$ is a step function in β_0 that makes incremental decreases with increasing β_0. Finally consider the permutation distribution $P = \sum_{i=1}^{n} q_i H_i$ whose distribution determines the mid-p-value as in (2.7). As noted in Theorem 1, the weights $\{q_i\}$ are $\{c_i\}$ for treatment survivals and $\{C_i\}$ for treatment censorings and these weights depend on the relative ordering of the censored treatment values with the control values. Thus the distribution of P changes with β_0 as a result of the changes in the weights that occur when a treatment

censoring interchanges with a control value. Thus the distribution of P is denoted as $P(\beta_0)$.

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

$$\hat{p}(\beta_0) = \Pr\{P(\beta_0) > v_0(\beta_0)\} + \tfrac{1}{2}\Pr\{P(\beta_0) = v_0(\beta_0)\}.$$

The set $\{\beta_0 : 0.025 \leq \hat{p}(\beta_0) \leq 0.975\}$ determines a 95% nominal interval.

The superior accuracy of the saddlepoint mid-p-value over the asymptotic normal theory suggests that inversion of the former should produce confidence intervals whose actual coverage is much closer to the nominal coverage. This is indeed the case. In implementing this inversion, the distribution of $P(\beta_0)$ and values of $v_0(\beta_0)$ are both step functions in β_0. Plots of $\hat{p}(\beta_0)$ should have the same step function appearance as the asymptotic normal theory methods used by Prentice.

As an example of intermediate size with light censoring, consider the vaginal cancer data of Pike (1966) given in Table 5.

Group 1	$143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 220,$
	$227, 230, 234, 248, 265, 304, 216^+, 244^+$
Group 2	$142, 156, 163, 198, 205, 232, 232, 233, 233, 233, 233,$
	$239, 240, 261, 280, 280, 296, 296, 323, 204^+, 344^+$

Table 5. Days to vaginal cancer in female rats from Pike (1966). $^+$Censored value.

Take $y = \log(t - 100)$ and a grid of β_0-values over $(-5, 5)$ with incremental step 0.001. For each β_0 value, the residuals $y - \beta_0 z$ were computed, ordered, and the normal and saddlepoint mid-p-values were computed for both the log-rank and generalized Wilcoxon statistics. For smaller values of β_0 used in determining the left edge of

the confidence interval, Figures 1 and 2 plot $\hat{p}(\beta_0)$ versus β_0 for the log-rank and generalized Wilcoxon statistics respectively. The figures provide the saddlepoint (dotted) and normal (solid) mid-p-values for the collection of one-sided tests of $H_0 : \beta = \beta_0$ versus $H_1 : \beta > \beta_0$ over the range $\beta_0 \in (-0.1,\ 0.055)$ with $\hat{p}(\beta_0) \in (0.0, 0.1)$.

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

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

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

Table 6. Confidence intervals for the Pike data.

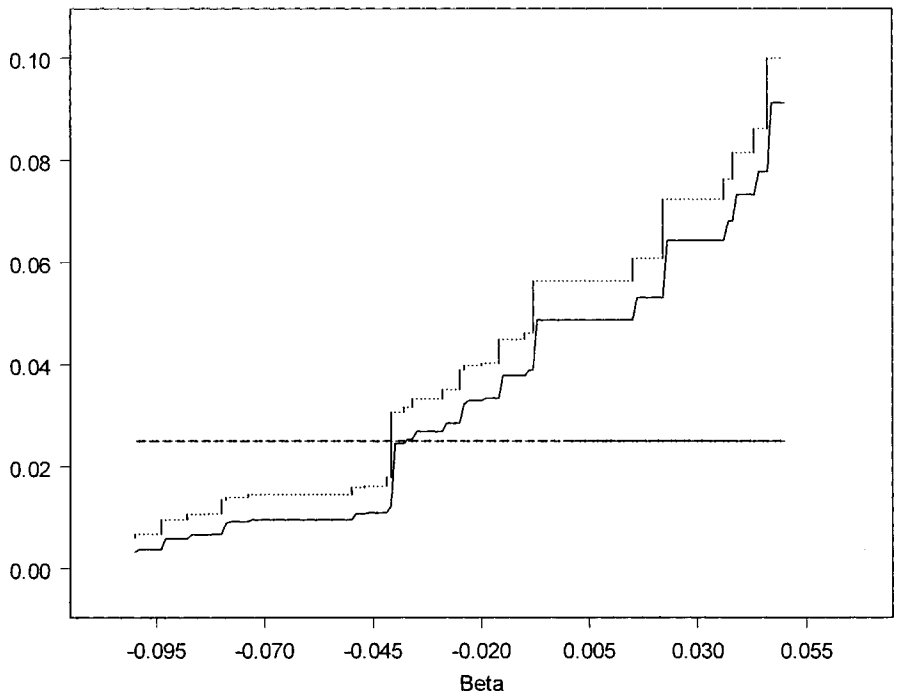

Figure 1. Plot of $\hat{p}(\beta_0)$ versus β_0 for the log-rank test using the Pike data. Mid-p-values for the saddlepoint approximations (dotted) and the normal approximations (solid) are shown over the range $(0.0,\ 0.1)$.

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

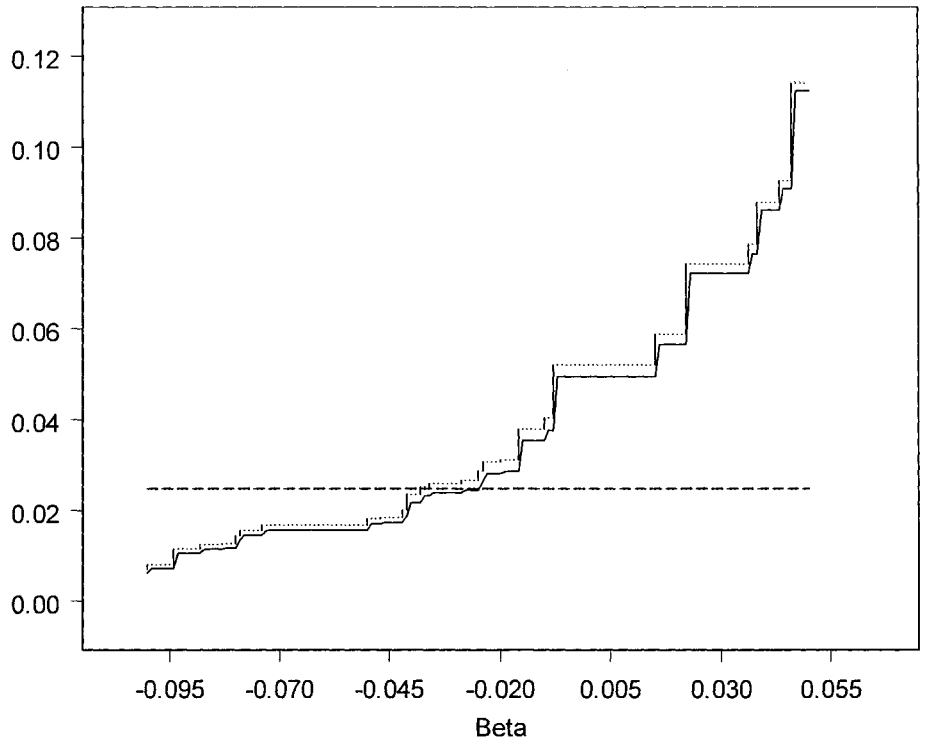

Figure 2. Same plot of $\hat{p}(\beta_0)$ versus β_0 as Figure 1 but for the generalized Wilcoxon test.

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

To determine which of the two confidence intervals are more accurate, the true confidence intervals are computed using the simulated mid-p-values as in §§2.2 and 2.3. Since it was prohibitive to simulate mid-p-values over the entire grid of β_0 values, only the true mid-p-values for a sequence of β_0 values that searched for a root to $p(\beta_0) = 0.025$ were used. Starting with the saddlepoint confidence interval,

this sequence of β_0-values consisted of those determined by using a bisection method to solve for the root of $p(\beta_0) = 0.025$.

For the Pike data, Table 6 summarizes the true confidence interval and the confidence intervals from the saddlepoint and normal approximations for both log-rank and generalized Wilcoxon tests. These intervals are all conservative because the determination of the endpoints was based on the vertical steps in the plot cutting through the horizontal line at height 0.025 and the coverage must include the full probability mass at the endpoints. It is clear that the saddlepoint confidence interval is more accurate than normal for both the log-rank and generalized Wilcoxon tests.

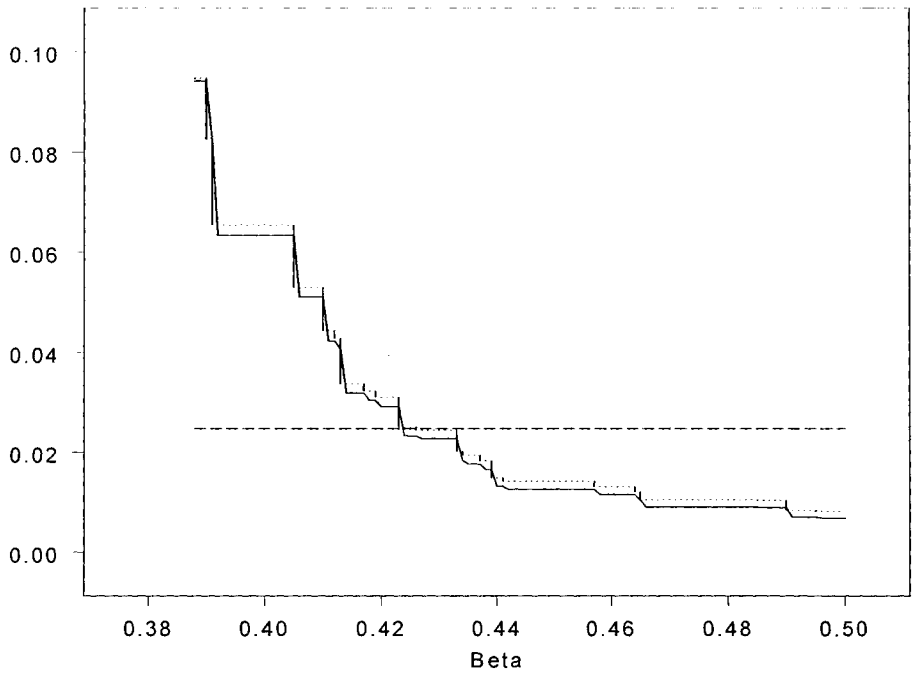

Figure 3. Plot of $1 - \hat{p}(\beta_0)$ versus β_0 for the log-rank test using the Pike data. Mid-p-values set against the alternative hypotheses $H_1 : \beta < \beta_0$ for the saddlepoint approximations (dotted) and the normal approximations (solid) are shown over the range $(0.0,\ 0.1)$.

Any attempt to simulate these confidence intervals, for the purpose of determining whether the saddlepoint interval achieves more accurate coverage, would be difficult to implement and also difficult to interpret because the true coverage cannot be set to 95%. As indicated above, even the true confidence interval using simulation is conservative due to the extra mass at the endpoints of the confidence interval that pushes the total coverage over 95%. However the simulations in §2.4 suggest that if

(a, b) is the true $100\alpha\%$ confidence interval with $\alpha > 0.95$, then $\hat{p}(a) \simeq p(a)$ and $\hat{p}(b) \simeq p(b)$ so that $\hat{p}(b) - \hat{p}(a) \simeq p(b) - p(a) = \alpha$ and the total coverage as well as overshoot and undershoot should all be very close to their true values.

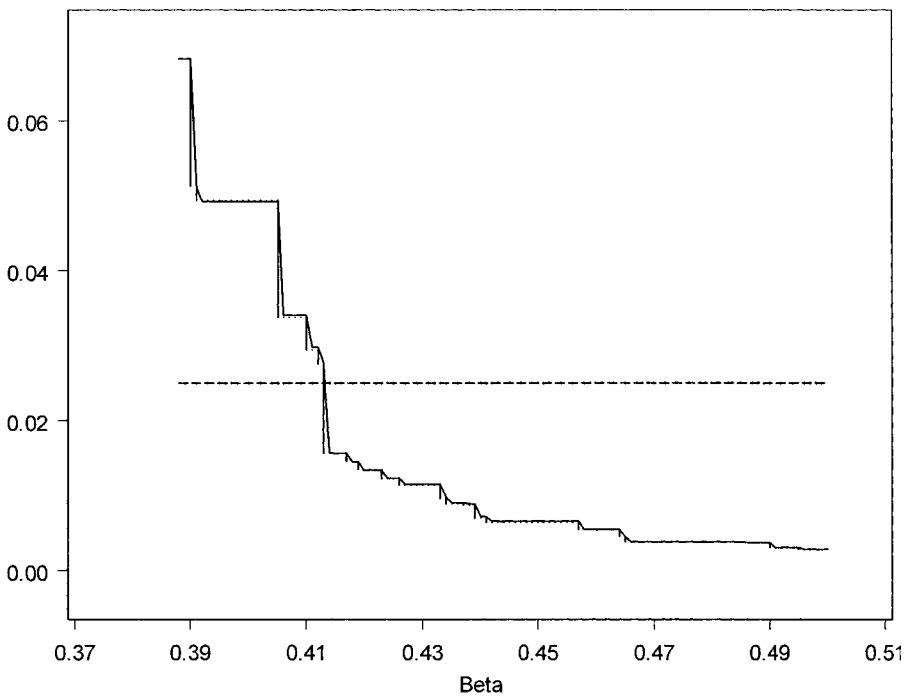

Figure 4. Same plot of $1 - \hat{p}(\beta_0)$ versus β_0 as Figure 3 but for the generalized Wilcoxon test.

2.3.1 Some Examples

As an alternative to the simulation of confidence intervals, a wide range of additional examples is considered. In each example, β is the differential effect attributed to the second treatment.

Breast Cancer: (Sedmak et. al., 1989)

This study is designed to determine if female breast cancer patients, originally classified as lymph node negative by standard light microscopy (SLM), could be more accurately classified by immunohistochemical (IH) examination of their lymph nodes with an anticytokeration, monoclonal antibody cocktail. The data for forty-five female breast-cancer patients with auxiliary negative lymph nodes and a minimum 10-year follow-up were selected from the Ohio State University Hospital's Cancer Registry. Of the 45 patients, 9 were immunoperoxidase positive, and the remaining 36 still remained negative. Survival times (in months) for both groups of patients are given in Table 15. This data is an example for unbalanced data sets.

IH Neg.	$19, 25, 30, 34, 37, 46, 47, 51, 56, 57, 61, 66, 67, 74, 78, 86, 122^+,$ $123^+, 130^+, 130^+, 133^+, 134^+, 136^+, 141^+, 143^+, 148^+, 151^+,$ $152^+, 153^+, 154^+, 156^+, 162^+, 164^+, 165^+, 182^+, 189^+$
IH Pos.	$22, 23, 38, 42, 73, 77, 89, 115, 144^+$

Table 7. Times to death for breast cancer patients. $^+$Censored.

The true, saddlepoint and normal confidence intervals of β for the log-rank and generalized Wilcoxon tests are shown at Table 8.

	Log-Rank		G. Wilcoxon	
	Lower	Upper	Lower	Upper
True	-2.14769	$-.19402$	-1.93874	.03498
Sad.	-2.113	$-.194$	-1.940	.035
Normal	-2.069	$-.278$	-1.953	.040

Table 8. Confidence intervals for β in the breast cancer data.

Chemotherapy in Ovarian Cancer: (Edmunson et al., 1979)

Surgical treatment of ovarian cancer may be followed by a course of chemotherapy. In a study of different chemotherapy treatments, Edmunson et. al. (1979) compared the antitumor effects of cyclophosphamide alone and cyclophosphamide combined with adriamycin. The trial involved 26 women. The response variable was the survival time in days. The data are given in Table 9. This data set is an example of a heavily censored data set.

Cycl.	$59, 115, 156, 268, 329, 431, 448^+, 477^+, 638, 803^+, 855^+,$ $1040^+, 1106^+$
Cycl./Adr.	$353, 365, 377^+, 421^+, 464, 475, 563, 744^+, 769^+, 770^+,$ $1129^+, 1206^+, 1227^+$

Table 9. Survival times of ovarian cancer patients. $^+$Censored

The true, saddlepoint and normal confidence intervals of β for log-rank and generalized Wilcoxon tests are shown at Table 10.

	Log-Rank		G.Wilcoxon	
	Lower	Upper	Lower	Upper
True	$-.7901$	3.0348	$-.5586$	2.9519
Sad	$-.808$	3.035	$-.559$	2.952
Normal	$-.676$	2.351	$-.527$	2.256

Table 10. Confidence intervals for β in the ovarian cancer data.

Myelomatosis: (Peto et al., 1977)

Table 11 gives survival times for 25 patients diagnosed with myelomatosis (Peto et al., 1977). These patients were randomly assigned to two drug treatments,

and the times in days from the point of randomization to either death or censoring are recorded. This data set is an example of a small size data set.

Treat. 1	$8, 8, 52, 63, 63, 220, 365^+, 852^+, 1296^+, 1328^+, 1460^+, 1976^+$
Treat. 2	$13, 18, 23, 70, 76, 180, 195, 210, 632, 700, 1296, 1990^+, 2240^+$

Table 11. Myelomatosis data.$^+$Censored.

The true, saddlepoint and normal confidence intervals of β for log-rank and generalized Wilcoxon tests are shown at Table 12.

	Log-Rank		G.Wilcoxon	
	Lower	Upper	Lower	Upper
True	-5.023	1.7739	-4.1829	2.1699
Sad	-4.722	1.774	-4.183	2.170
Normal	-4.056	1.774	-3.335	2.170

Table 12. Confidence intervals for β in the myelomatosis data.

Gastric Carcinoma: (Stablein et al., 1981)

Table 13 provides survival times in days from a clinical trial on gastric carcinoma that involves 90 patients randomized to either chemotherapy alone or to a combination of chemotherapy and radiation. This is an example of a large data set with light censoring.

chem.	$17, 42, 44, 48, 60, 72, 74, 95, 103, 108, 122, 144, 167, 170, 183,$ $185, 193, 195, 197, 197, 208, 234, 235, 254, 307, 315, 401, 445,$ $464, 484, 528, 542, 567, 577, 580, 795, 855, 882^+, 882^+, 892^+,$ $1031^+\ 1033^+, 1306^+, 1335^+, 1366, 1452^+, 1472^+$
chem./rad.	$1, 63, 105, 129, 182, 216, 250, 262, 301, 301, 301, 342, 354,$ $356, 358, 380, 381^+, 383, 383, 388, 394, 408, 460, 489, 499,$ $524, 529^+, 535, 535, 562, 562, 675, 676, 748, 748, 778, 786,$ $797, 945^+, 955, 968, 1180^+, 1245, 1271, 1277^+, 1397^+, 1512^+,$ 1519^+

Table 13. Survival times for gastric carcinoma. $^{+}$Censored.

The true, saddlepoint and normal confidence intervals of β for log-rank and generalized Wilcoxon tests are shown in Table 14.

	Log-Rank		G.Wilcoxon	
	Lower	Upper	Lower	Upper
True	$-.2575$	.9099	.0685	1.0039
Sad	$-.258$	.910	.068	1.004
Nor	$-.229$	.885	.081	1.000

Table 14. Confidence intervals for β in the gastric carcinoma example.

2.3.2 Conclusions

The simulation in §2.3 and the previous examples suggest that the saddlepoint approximation offers an accurate approach for estimating the true confidence intervals of the location parameter β for the widely used log-rank score test and also for the generalized Wilcoxon score test in survival data analysis. Whatever the data type, the saddlepoint approximation is extremely accurate and consistently more accurate than the normal approximation. The saddlepoint approximation is recommended for use with all data sets regardless of the size of the data set, the degree of imbalance and the amount of censoring.

2.4 Tied Values

Modifications to the presentation above are needed to deal with tied survival times involving members from different groups. When the ties involve members from the

same group, e.g. the same z, then the rank tests are invariant to the ordering of the ties and so ties can be ignored by assuming any particular ordering.

Two strategies are generally employed to deal with ties. First, tied values can be assigned the average of their possible scores. This is the easier way to deal with the issue since only one mid-p-value computation is needed. Secondly, and perhaps more appropriately but also more difficult computationally, the rank vector can be allowed to assume all possible underlying rank vectors that would be consistent with the observed ties and censored data (Kalbfleisch and Prentice, 2002, p. 234). In this approach, the overall mid-p-value is the average of all mid-p-values computed over the set of consistent rank vector assignments.

In this section we show how to use these two methods to calculate the mid-p-value for some tied valued examples. Some points help when considering these calculations:

1. Tied censored values have the same scores for log-rank and generalized Wilcoxon regardless of group membership, so no changes are needed for tied censored values.

2. When the tied values involve survival times and censoring times, it is natural to arrange the censored time after all the survival times by shifting the censoring time by a small amount like 10^{-6}.

3. If ties belong to the same group, then the particular ordering does not matter.

4. Only ties among survival times from different groups need to be dealt with.

2.4.1 Examples

For illustration consider the following examples.

Survival data for Chronic Active Hepatitis patients : (Kirk, A.P., et al., 1980)

These data are survival times (in months) for patients suffering from chronic active hepatitis. Times are given for a control group and a group being treated with prednisolone.

Control	$2, 3, 4, 7, 10, 22, 28, 29, 32, 37, 40, 41, 54, 61, 63, 71, 127^+,$ $140^+, 146^+, 158^+, 167^+, 182^+$
Prednisolone	$2, 6, 12, 54, 56^+, 68, 89, 96, 96, 125^+, 128^+, 131^+, 140^+,$ $141^+, 143, 145^+, 146, 148^+, 162^+, 168, 173^+, 181^+$

Table 15. Survival data for chronic active hepatitis patients. $^+$Censored.

Note this data set contains 5 tied times but according to the list of points, the actual ties that affect the calculation are the two at the times 2 and 54. The tied points at 96 are from the same group, the tied points at 140 are both censored, and the tied points at 146 are a death and censored.

Death times of Psychiatric Patients: (Woolson, R.F., 1981)

Survival times (in months) for 26 psychiatric inpatients admitted to the University of Iowa Hospital during the years 1935-1948 are recorded in Table 16. The data are 15 female and 11 male patients.

Female	$1, 1, 2, 11, 14, 22, 24, 26, 31^+, 32, 35^+, 35^+, 36^+, 37^+, 40$
Male	$22, 25, 28, 30^+, 30^+, 31^+, 33^+, 33^+, 34^+, 35, 39^+$

Table 16. Death times of Psychiatric Patients. $^{+}$Censored

This data set has 5 tied times but only the ties at time 22 are relevant.

Autologous and Allogeneic Bone Marrow Transplants: (Klein, J.P., Moechberger, M.L., 1997)

Data in Table 17 are a sample of 101 patients with advanced acute myelogenous leukemia reported to the international bone marrow transplant registry. Fifty-one received an autologous (Auto) bone marrow transplant and fifty patients had an allogeneic (Allo) bone marrow transplant. Interest is in comparing the effectiveness of these two methods of transplant.

Allo	$0.030, 0.493, 0.855, 1.184, 1.283, 1.480, 1.776, 2.138, 2.500, 2.763, 2.993,$ $3.224, 3.421, 4.178, 4.441^{+}, 5.691, 5.8551^{+}, 6.941, 6.941^{+}, 7.993^{+}, 8.882,$ $8.882, 9.145^{+}, 11.480, 11.51312.105^{+}, 12.796, 12.993^{+}, 13.849^{+}, 16.612^{+},$ $17.138^{+}, 20.066, 20.329^{+}, 22.368^{+}, 26.776^{+}, 28.717^{+}, 28.717^{+}, 32.928^{+},$ $33.783^{+}, 34.221^{+}, 34.770^{+}, 39.593^{+}, 41.118^{+}, 45.003^{+}, 46.053^{+}, 46.941^{+},$ $48.289^{+}, 57.401^{+}, 58.322^{+}, 60.625^{+}$
Auto	$0.658, 0.822, 1.414, 2.500, 3.322, 3.816, 4.737, 4.836^{+}, 4.934, 5.033, 5.757,$ $5.855, 5.987, 6.151, 6.217, 6.447^{+}, 8.651, 8.711, 9.441^{+}, 10.329, 11.480,$ $12.007, 12.007^{+}, 12.237, 12.401^{+}, 13.059^{+}, 14.474^{+}, 15.000^{+}, 15.461,$ $15.757, 16.480, 16.711, 17.204^{+}, 17.237, 17.303^{+}, 17.664^{+}, 18.092,$ $18.092^{+}, 18.750^{+}, 20.625^{+}, 23.158, 27.730^{+}, 31.184^{+}, 32.434^{+}, 35.921^{+},$ $42.237^{+}, 44.638^{+}, 46.480^{+}, 47.467^{+}, 48.322^{+}, 56.086$

Table 17. Leukemia free survival times (in months) for Auto and Allo Transplants.$^{+}$Censored

This data sets has 7 tied times but only the two ties occurring at times 2.5 and 11.48 need be considered.

2.4.2 Score Average Method

Tied values can be dealt with by assigning the average of the tied scores to all the ties occurring at the same time. From (2.3) and (2.5), it is clear that the score average method requires only a single computation of mid-p-value. Table 18, shows the true, saddlepoint and normal mid-p-values for log-rank and generalized Wilcoxon tests for the above examples using the score average method.

Data Set	Log-Rank			G.Wilcoxon		
	True	Sadpt.	Normal	True	Sadpt.	Normal
Hepatitis	.0176	.0178	.0158	.0077	.0077	.0070
Psychiatric	.0971	.1007	.1022	.0535	.0540	.0555
Auto/Allo	.2703	.2700	.2694	.5005	.4999	.4997

Table 18. Mid-p-values using score averages for ties in the three examples.

As expected, the saddlepoint approximation maintains its accuracy with score average method and is more accurate in all three of the examples.

2.4.3 Permutation Method

This method is the more natural way to deal with ties, since all possible permutations of the tied values have the same probability of occurrence. The average mid-p-value is the proper method for combining the set of mid-p-values. Only permutations of tied deaths from different groups need to be considered. Table 19, reports the true, saddlepoint and normal mid-p-values for log-rank and generalized Wilcoxon tests for the above examples using this permutation method.

Data Set	Log-Rank			G.Wilcoxon		
	True	Sadpt.	Normal	True	Sadpt.	Normal
Hepatitis	.0177	.0178	.0159	$.0^{2}780$	$.0^{2}779$	$.0^{2}700$
Psychiatric	.0910	.0957	.0971	.0480	.0480	.0494
Auto/Allo	.2704	.270	.2694	.4986	.4983	.4982

Table 19. Mid-p-values using the permutation method for ties in the three examples.

Using the permutation method, the saddlepoint approximation again shows superior accuracy.

2.4.4 Conclusion

The saddlepoint approximation has succeeded in estimating the true mid-p-value with deviations of less than 0.001. From Tables 18 and 19, it is apparent that the score average provides good estimates for the mid-p-value for large sample sizes. This occurs because the changes in the scores for the treatment group when using the permutation method are very small with large sample sizes and lead to averages which are roughly the score average mid-p-value. For small samples however, the permutation method may be better. For both methods, greater accuracy results by using the saddlepoint approximation.

Chapter 3
Saddlepoint Tests for the Log-Rank and Generalized Wilcoxon Type Trend Tests

3.1 Introduction

Since the 1950's there has been considerable interest in procedures for ordered location alternatives in multisample data. Ordered alternatives refer to a generalization of one sided alternatives in the two sample problem: If $\{\beta_i : i = 1, ..., p+1\}$ represent location parameters for $p+1$ populations, an order alternative is one that specifies a particular ordering of the β_i's, prior to observation of the data. One of the earliest works in this area was by Jonckheere (1954) in the one-way layouts based on Kendall's test for rank correlation. Another more convenient and popular form was given by Page (1963). For censored data, the generalization of Wilcoxon type trend tests are the most widely used in clinical trials. In this chapter we present saddlepoint approximations for the permutation distributions of the log-rank trend test and the Peto-Prentice generalization of the Wilcoxon trend test.

3.2 Log-Rank and Generalized Wilcoxon Type Trend Tests

3.2.1 Log-Rank Test for Trend

In order to motivate the tests for trend, the original development of the two sample log-rank test is first presented. Generalization to $(p+1)$ samples then leads to the test for trend. For the two sample test, suppose that there are k distinct death times $t_{(1)} < t_{(2)} < ... < t_{(k)}$, and that at time $t_{(j)}$, d_{1j} individuals in group 1 and d_{2j} individuals in group 2 die, for $j = 1, ..., k$. Suppose further that there are n_{1j} individuals at risk in group 1 at time $t_{(j)}^-$, and there are n_{2j} at risk in group 2, consequently, $d_j = d_{1j} + d_{2j}$ deaths in total out of $n_j = n_{1j} + n_{2j}$ individuals at risk. At each $t_{(j)}$ we have a 2×2 contingency table as in Table 1.

	death	survive	
group1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
group2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
	d_j	$n_j - d_j$	

Table 1: Two sample 2×2 contingency table

Conditional on the row and column total as fixed, d_{1j} has a hypergeometric distribution with mean $e_{1j} = n_{1j}d_j/n_j$. In testing the null hypothesis that there is no difference in the survival experiences in the two groups, Mantel-Hanszel (1959) proposed the statistic

$$v_{MH} = \sum_{j=1}^{k}(d_{1j} - e_{1j}). \tag{3.1}$$

If the contingency tables are independent, the variance of v_{MH} is

$$Var(v_{MH}) = \sum_{j=1}^{k} \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

and asymptotically normal with $v_\ell / \sqrt{Var(v_\ell)} \sim N(0,1)$. This derivation of the log-rank test is similar to that given by Mantel (1966). It is difficult, however, to formalize the distribution theory from this development since the contingency tables over failure times are clearly not independent. It can, however, be shown that the terms $(d_{1j} - e_{1j})$ are uncorrelated and that $Var(v_\ell)$ provides an estimate of the variance, see Kalbfleisch and Prentice (2002) section 1.5.

For the $p+1$ sample test, an obvious extension of this notation can be used to give

$$v_{\ell i} = \sum_{j=1}^{k}(d_{ij} - \frac{n_{ij}d_j}{n_j}) = (d_{i.} - e_{i.}), \tag{3.2}$$

as the log-rank discrepancy in deaths for group $i = 1, ..., p+1$. These quantities are then expressed in the $p \times 1$ vector form $v_\ell^- = (v_{\ell 1}, ..., v_{\ell p})^T$, and the $(p+1) \times 1$ vector $v_\ell = (v_{\ell 1}, ..., v_{\ell,p+1})^T$. Then v_ℓ^- has a variance-covariance matrix whose $(i,j)^{th}$ component is

$$\sum_{l=1}^{k} \frac{n_{il}d_l(n_l - d_l)}{n_l^2(n_l - 1)}(\delta_{ij} - \frac{n_{il}}{n_l}), \qquad i,j = 1, ..., p$$

where δ_{ij} is the indicator that $i = j$. In order to test the null hypothesis of no group difference, the statistic $(v_\ell^-)^T\{Var(v_\ell^-)\}^{-1}v_\ell^-$ is compared to a χ_p^2 distribution; see Kalbfleisch and Prentice (2002, §1.5).

Consider now the test for trend where $p+1$ groups of survival data are to be compared, and where these groups are ordered in some way. For example, the

groups may correspond to increasing doses of a treatment, the stage of a disease, or the age-group of an individual, and we want to test

$$H_0 : S_1(t) = S_2(t) = ... = S_{p+1}(t)$$

versus

$$H_1 : S_1(t) \leq S_2(t) \leq ... \leq S_{p+1}(t)$$

with at least one of the equalities in H_1 a strict inequality. The log-rank test for trend across $p+1$ ordered groups is based on the statistic

$$u_\ell = l^T v_\ell = \sum_{i=1}^{p+1} l_i(d_{i.} - e_{i.}) \tag{3.3}$$

where the components of $l = (l_1, ..., l_{p+1})^T$ are the group codes or doses that are constants. The p-value for testing H_0 uses the asymptotic normality of $u_\ell / \sqrt{Var(u_\ell)}$, see Collett (2003).

3.2.2 Generalized Wilcoxon-Type Trend Test:

Similar to the $p+1$ log-rank sample test, the so called Peto-Prentice generalization for the Wilcoxon test can be written as an weighted form of (3.1). Let $v_\omega = (v_{\omega 1}, ...v_{\omega,p+1})^T$ where

$$v_{\omega i} = \pi_i(d_{i.} - e_{i.}), \qquad \pi_i = \prod_{j=1}^{i}(1 - \frac{d_{ij}}{n_j + 1}) \tag{3.4}$$

for $i = 1, ..., p+1$. The generalized Wilcoxon trend test uses the statistic

$$u_\omega = l^T v_\omega = \sum_{i=1}^{p+1} l_i v_{\omega i}.$$

3.3 Accelerated Failure Time Model Setting

The AFT model was explained in chapter 2 as the model

$$y_i = \alpha + \beta^T z_i + \sigma e_i, \qquad i = 1, ..., n \tag{3.5}$$

where $y_i = \ln t_i$, $\beta = (\beta_1, ..., \beta_{p+1})$ is a vector of coefficients, and z_i is a $(p+1)$ covariate vector. The test of hypothesis that there is a trend among groups is specified as.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p+1}$$

versus

$$H_1 : \beta_1 \leq \beta_2 \leq ... \leq \beta_{p+1} \tag{3.6}$$

with at least one strict inequality in the alternative. Vector z_i is a $(p+1)$ vector that indicates the group membership of y_i; e.g. it has a 1 in the j^{th} component and zeros elsewhere if y_i belongs to group j. Let $\beta_i = \beta l_i$, so that the test in (3.6) becomes $H_0 : \beta = 0$ vs $H_1 : \beta > 0$, where $l^T = (l_1, ..., l_{p+1})$ is vector of constants (doses), which may be either equally spaced or not. The model (3.5) now becomes

$$y_i = \alpha + \beta l^T z_i + \sigma e_i, \qquad i = 1, ..., n$$

with error probability density function $f(e)$. With the notation change $l^T z_i = x_i$, then

$$y_i = \alpha + \beta x_i + \sigma e_i, \qquad i = 1, ..., n$$

is a special case of the AFT model (3.5) in which z_i is one dimensional. Thus the same results of Prentice (1978), using the marginal density of the rank vectors, give

the score test statistic for the trend test as

$$\begin{aligned} u &= \sum_{i=1}^{k}\left(c_i x_{(i)} + C_i \sum_{j=1}^{m_i} x_{ij}\right) \\ &= l^T\left[\sum_{i=1}^{k}\left(c_i z_{(i)} + C_i \sum_{j=1}^{m_i} z_{ij}\right)\right] \end{aligned} \tag{3.7}$$

where k is the number of distinct death times and $\{c_i\}$ and $\{C_i\}$ are constants determined using the specific error distribution. In particular, using the logistic error distribution gives minus the generalized Wilcoxon trend test while using the extreme value distribution leads to minus the log-rank trend test explained in section 2. To see this, consider the $(p+1) \times 1$ statistic

$$v = \sum_{i=1}^{k}\left(c_i z_{(i)} + C_i \sum_{j=1}^{m_i} z_{ij}\right) \tag{3.8}$$

where $z_{(1)}, z_{(2)}, ..., z_{(k)}$ are $(p+1) \times 1$ vectors, and $z_{(i)}$ is the group membership vector for the individual dying at log time $y_{(i)}$, and likewise for $z_{i1}, ..., z_{im_i}$ for the m_i censored individuals in $[y_{(i)}, y_{(i+1)})$ for $i = 1, ..., k$ with $y_{(0)} = -\infty$, $y_{(k+1)} = \infty$. With no ties, the vector $v_\ell = (v_{\ell 1}, ..., v_{\ell p+1})^T$ statistic can be written as

$$v_\ell = \sum_{j=1}^{k}(z_{(j)} - \bar{z}_{(j)}), \qquad \bar{z}_{(j)} = n_j^{-1} \sum_{l \in R(t_{(j)})} z_l \tag{3.9}$$

where $R(t_{(j)})$ denotes those individuals under risk at time $t_{(j)}^-$.

Lemma *The statistic v_ℓ in (3.9) is the log-rank score vector, or minus the censored generalization of the Savage exponential score vector represented by Prentice (1978) in the linear rank form (3.8).*

Proof

$$
\begin{aligned}
v_\ell &= \sum_{j=1}^{k}\left(z_{(j)} - n_j^{-1}\sum_{l\in R(t_{(j)})} z_l\right)\\
&= z_{(1)} - n_1^{-1}\left(z_{(1)} + \sum_{j=1}^{m_1} z_{1j} + z_{(2)} + \sum_{j=1}^{m_2} z_{2j} + \ldots\right)\\
&\quad + z_{(2)} - n_2^{-1}\left(z_{(2)} + \sum_{j=1}^{m_2} z_{2j} + z_{(3)} + \sum_{j=1}^{m_3} z_{3j} + \ldots\right) + \ldots\\
&\quad + z_{(k)} - n_k^{-1}\left(z_{(k)} + \sum_{j=1}^{m_k} z_{kj}\right)
\end{aligned}
$$

Rearrangement of the sum gives

$$
\begin{aligned}
v_\ell &= \left(z_{(1)} - n_1^{-1}z_{(1)}\right) - n_1^{-1}\sum_{j=1}^{m_1} z_{1j}\\
&\quad + \left(z_{(2)} - n_1^{-1}z_{(2)} - n_2^{-1}z_{(2)}\right) - \left(n_1^{-1} + n_2^{-1}\right)\sum_{j=1}^{m_2} z_{2j} + \ldots\\
&\quad + \left\{z_{(k)} - \left(\sum_{i=1}^{k} n_i^{-1}\right) z_{(k)}\right\} - \left(\sum_{i=1}^{k} n_i^{-1}\right)\sum_{j=1}^{m_k} z_{kj}\\
&= -\sum_{j=1}^{k}\left\{\left(\sum_{i=1}^{j} n_i^{-1} - 1\right) z_{(j)} + \left(\sum_{i=1}^{j} n_i^{-1}\right)\sum_{i=1}^{m_j} z_{ji}\right\}\\
&= -v
\end{aligned}
$$

which is the value in (3.8).■

Here the weights

$$c_j = \sum_{i=1}^{j} n_i^{-1} - 1, \qquad C_j = \sum_{i=1}^{j} n_i^{-1}$$

are the same as in Chapter 2.

Prentice (1978) used the asymptotically normality of $v_\ell^{-T}(V_0^-)^{-1}v_\ell^-$ to estimate the p-value, where $v_\ell^- = (v_{\ell 1}, ..., v_{\ell p})^T$ and V_0^- is the first $(p \times p)$ components of

$$V_0 = Var(v_\ell) = \sum_{i=1}^{k} n_i^{-1} \sum_{l \in R\{t_{(i)}\}} (z_l - \bar{z}_i)(z_l - \bar{z}_i)^T, \qquad \bar{z}_i = n_i^{-1} \sum_{l \in R\{t_{(i)}\}} z_l,$$

Lemma *For the generalized Wilcoxon, with no ties, the* $(p+1) \times 1$ *vector in (3.4) can be written as*

$$v_\omega = \sum_{i=1}^{k} \pi_i (z_{(i)} - \bar{z}_{(i)}), \qquad \pi_i = \prod_{j=1}^{i} \frac{n_j}{n_j + 1} \tag{3.10}$$

which, in the linear rank form (3.8), is minus the generalized Wilcoxon represented in Chapter 2 with

$$c_i = 1 - 2\prod_{j=1}^{i} \frac{n_j}{n_j + 1}, \qquad C_i = 1 - \prod_{j=1}^{i} \frac{n_j}{n_j + 1}$$

Proof

$$\begin{aligned} v_\omega &= -\sum_{i=1}^{k} \pi_i (z_{(i)} - \bar{z}_{(i)}) \\ &= \sum_{i=1}^{k} (c_i - C_i)(z_{(i)} - \bar{z}_{(i)}) \end{aligned}$$

but

$$c_i - C_i = -n_i (C_i - C_{i-1}) \tag{3.11}$$

since

$$\begin{aligned} -n_i(C_i - C_{i-1}) &= -n_i \{1 - \pi_i - (1 - \pi_{i-1})\} \\ &= -n_i(-\pi_i + \frac{n_i + 1}{n_i}\pi_i) \\ &= -\pi_i = c_i - C_i \end{aligned}$$

then

$$\begin{aligned} v_\omega &= \sum_{i=1}^{k} \left\{ (c_i - C_i) z_{(i)} + n_i (C_i - C_{i-1}) \bar{z}_{(i)} \right\} \\ &= \sum_{i=1}^{k} \left\{ c_i z_{(i)} + C_i \left(n_i \bar{z}_{(i)} - z_{(i)} - n_{i+1} \bar{z}_{(i+1)} \right) \right\} \\ &= \sum_{i=1}^{k} \left(c_i z_{(i)} + C_i \sum_{j=1}^{m_i} z_{ij} \right) \end{aligned}$$

For the proof of the general case of (3.11) see Kalbfleisch and Prentice (2002) , Mehrotra et al., (1982) and Andersen et al., (1982).■

The normal p-value for this test makes use of the asymptotically normal of $v_\omega^{-T}(V_0^-)^{-1}v_\omega^-$, where V_0^- is the first $(p \times p)$ components of

$$\begin{aligned} Var(v_\omega) &= \sum_{i=1}^{k} [a_i(1 - a_i^*)\{2z_{(i)}z_{(i)}^T + \sum_{j=1}^{m_i} z_{ij}z_{ij}^T\} \\ &\quad -(a_i^* - a_i)x_{(i)}\{a_i x_{(i)}^T + 2\sum_{j=i+1}^{k} a_j x_{(j)}^T\}] \end{aligned}$$

and

$$a_i = \prod_{j=1}^{i} \frac{n_j}{n_j + 1}, \qquad a_i^* = \prod_{j=1}^{i} \frac{n_j + 1}{n_j + 2}, \qquad x_{(i)} = 2z_{(i)} + S_{(i)}, \qquad S_{(i)} = \sum_{j=1}^{m_i} z_{ij}.$$

The linear rank form (3.8) is simply $\sum_{i=1}^{n} q_i z_i$ so the trend test statistics u_ℓ and u_ω can be written as

$$l^T \sum_{i=1}^{n} q_i z_i.$$

The linearity of these statistics in the $\{z_i\}$ enables us to use the saddlepoint approximation to calculate the p-value for the test of the ordered alternative as shown in the next section.

3.4 Saddlepoint Approximation

Consider the permutation distribution for the statistic

$$u = l^T \sum_{i=1}^{n} q_i z_i \tag{3.12}$$

where $z_1, ..., z_n$ are $(p+1) \times 1$ group membership vectors. The vectors $z_1, ..., z_n$ are assumed to be random with a uniform distribution over all one way designs for $(p+1) \times 1$ vectors such that $\sum_{i=1}^{n} z_i = (n_1, ..., n_{p+1})^T$. Such a distribution assigns probability

$$\begin{pmatrix} n \\ n_1, ..., n_{p+1} \end{pmatrix}^{-1}$$

to each $z_1, ..., z_n$ outcome. This distribution is also the conditional mass function

$$z_1, ..., z_n \stackrel{D}{=} \zeta_1, ..., \zeta_n | \sum_{i=1}^{n} \zeta_i = (n_1, ..., n_{p+1})^T$$

where $\zeta_1, ..., \zeta_n$ are i.i.d. Multinomial$(1, \theta_1, ..., \theta_{p+1})$. To prevent a degenerate mass function, the $p+1$ components of $z_1, ..., z_n$ are dropped to give $p \times 1$ vectors $z_1^-, ..., z_n^-$. The resulting non-degenerate distribution is

$$z_1^-, ..., z_n^- \stackrel{D}{=} \zeta_1^-, ..., \zeta_n^- | \sum_{i=1}^{n} \zeta_i^- = (n_1, ..., n_p)^T$$

where ζ_i^- also consists of the first p components in ζ_i. Now u can be written as a function of $z_1^-, ..., z_n^-$, say $u(z^-)$, and

$$\Pr\{u(z^-) \geq u_0\} = \Pr\left\{ u(\zeta^-) \geq u_0 | \sum_{i=1}^{n} \zeta_i^- = (n_1, ..., n_p)^T \right\}.$$

In order to write u as $u(z^-)$ note that,

$$\begin{aligned} u &= l^T \sum_{i=1}^{n} q_i z_i \\ &= l_p^T \sum_{i=1}^{n} q_i z_i^- + l_{p+1} \sum_{i=1}^{n} q_i z_{i,p+1} \\ &= \sum_{i=1}^{n} q_i \left(\sum_{j=1}^{p} l_j z_{ij} \right) + l_{p+1} \sum_{i=1}^{n} q_i \left(1 - \sum_{j=1}^{p} z_{ij} \right) \\ &= \sum_{i=1}^{n} q_i \left\{ \sum_{j=1}^{p} (l_j - l_{p+1}) z_{ij} \right\} + l_{p+1} \sum_{i=1}^{n} q_i \\ &= l_{p-}^T \sum_{i=1}^{n} q_i z_i^- + Q \end{aligned} \tag{3.13}$$

where $l_p^T = (l_1, ..., l_p)$, $l_{p-}^T = (l_1 - l_{p+1}, ..., l_p - l_{p+1})$ and $Q = l_{p+1} \sum_{i=1}^{n} q_i = 0$, since $\sum_{i=1}^{n} q_i = \sum_{i=1}^{k} (c_i + m_i C_i) = 0$, and m_i is the number of censored individuals in $[y_{(i)}, y_{(i+1)})$. See Kalbfleisch and Prentice (2002, eqn. 7.20).

Theorem 2 *Suppose that* $\zeta_1^-, ..., \zeta_n^-$ *are i.i.d. Multinomial* $(1, \theta_1, ..., \theta_p)$ *for any allowable* $\{\theta_i\}$, *e.g. such that* $\theta_i > 0$, $i = 1, ..., p$ *and* $\sum_{i=1}^{p} \theta_i < 1$. *Then the conditional distribution of* $\zeta^- = (\zeta_1^-, ..., \zeta_n^-)^T$ *given* $\sum_{i=1}^{n} \zeta_i^- = (n_1, ..., n_p)^T$ *is the marginal permutation distribution for* z^-. *This provides a conditional characterization for the null distribution of*

$$u = l_{p-}^T \sum_{i=1}^{n} q_i z_i^-$$

as

$$u \sim l_{p-}^T \sum_{i=1}^{n} q_i \zeta_i^- \quad \textit{given} \quad \sum_{i=1}^{n} \zeta_i^- = (n_1, ..., n_p)^T.$$

The weights $\{q_i\}$ are values from $\{c_i\}$ when associated with survival times and values from $\{C_i\}$ when associated with censored times. In the application, weight q_i depend on the relative placement of the censored values in the ordered sample.

The Skovgaard (1987) saddlepoint approximation leads to a conditional characterization as in Theorem 1. Let

$$Y = l_{p-}^T \sum_{i=1}^{n} q_i \zeta_i^- \qquad and \qquad X = \sum_{i=1}^{n} \zeta_i^- .$$

The Skovgaard approximation is based on the joint MGF of X and Y, derived as follow

$$\begin{aligned} M_{X,Y}(s,t) &= E\left\{\exp\left(t l_{p-}^T \sum_{i=1}^{n} q_i \zeta_i^- + s^T \sum_{i=1}^{n} \zeta_i^-\right)\right\} \\ &= \prod_{i=1}^{n} E\left\{\exp\left(t q_i l_{p-}^T + s^T\right)\zeta_i^-\right\} \\ &= \prod_{i=1}^{n}\left\{\left[\sum_{j=1}^{p} \theta_j \exp\left\{q_i(l_j - l_{p+1})t + s_j\right\}\right] + \theta_{p+1}\right\} \end{aligned}$$

where $s^T = (s_1, ..., s_p)^T$. Now let $r_{ij} = q_i(l_j - l_{p+1})$, then

$$M_{X,Y}(s,t) = \prod_{i=1}^{n}\left\{\left[\sum_{j=1}^{p} \theta_j \exp\left(r_{ij}t + s_j\right)\right] + \theta_{p+1}\right\} \tag{3.14}$$

Let $K(s,t) = \log M_{X,Y}(s,t)$ be the joint cumulant generating function (CGF). If u_0 is the observed value of the statistic (3.12), the p-value from the Skovgaard approximation is

$$\Pr(u \geq u_0) = \Pr(Y \geq u_0 \,|X = x) \simeq 1 - \Phi(\hat{w}) - \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right) \tag{3.15}$$

where $x = (n_1, ..., n_p)^T$, and

$$\hat{w} = \text{sgn}(\hat{t})\sqrt{2\left[\{K(\hat{s}_0, 0) - \hat{s}_0^T x\} - \{K(\hat{s}, \hat{t}) - \hat{s}^T x - u_0 \hat{t}\}\right]}$$
$$\hat{u} = \hat{t}\sqrt{|K''(\hat{s}, \hat{t})|/|K''_{ss}(\hat{s}_0, 0)|}.$$

In these expressions, K'' is the $(p+1) \times (p+1)$ Hessian matrix and K''_{ss} is the $\partial^2/\partial s_i \partial s_j$ $(p \times p)$ part at $(\hat{s}_0, 0)$. The numerator saddlepoint $(\hat{s}, \hat{t})$ with $s^T = (s_1, ..., s_p)^T$ solves

$$K'_{s_l}(\hat{s}, \hat{t}) = \sum_{i=1}^{n} \frac{\theta_l \exp(\hat{s}_l + r_{il}\hat{t})}{\left\{\sum_{j=1}^{p} \theta_j \exp(s_j + r_{ij}t) + \theta_{p+1}\right\}} = n_l, \qquad l = 1, ..., p$$
$$K'_t(\hat{s}, \hat{t}) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{p} \theta_j r_{ij} \exp(s_j + r_{ij}t)}{\left\{\sum_{j=1}^{p} \theta_j \exp(s_j + r_{ij}t) + \theta_{p+1}\right\}} = u_0.$$

Using $\theta_j = n_j/n$

$$K'_{s_l}(\hat{s}, \hat{t}) = \sum_{i=1}^{n} \frac{n_l \exp(\hat{s}_l + r_{il}\hat{t})}{\left\{\sum_{j=1}^{p} n_j \exp(s_j + r_{ij}t) + n_{p+1}\right\}} = n_l, \qquad l = 1, ..., p$$
$$K'_t(\hat{s}, \hat{t}) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{p} n_j r_{ij} \exp(s_j + r_{ij}t)}{\left\{\sum_{j=1}^{p} n_j \exp(s_j + r_{ij}t) + n_{p+1}\right\}} = u_0$$

and the denominator saddlepoint $\hat{s}_0$ solves

$$K'_{s_l}(\hat{s}_0, 0) = \frac{n \exp(s_l)}{\sum_{j=1}^{p} n_j \exp(s_j) + n_{p+1}} = 1$$

which gives $\hat{s}_0 = 0$.

3.5 Examples

Survival times for melanoma patients

The following data set is part of a study carried out by the University of Oklahoma Health Sciences center. The malignant tumor was surgically removed before allocation to Bacillus Calmette-Guerin(BCG) vaccine for three age groups, 21-40, 41-60 and 61-, see Lee, E.T.(1992). The survival times for the patients are

Group	$\{l_i\}$	$\{n_i\}$	Survival Time
21-40	-1	6	$19, 24^*, 8, 17^*, 17^*, 34^*$
41-60	0	3	$34^*, 4, 17^*$
61-	1	2	$10, 5$

with doses $\{l_i\} = \{-1, 0, 1\}$. Table 2 shows the true, normal and saddlepoint p-values.

Test stat.	Normal p	Sadpt. mid-p	True mid-p
LR	.048327	.067772	.071949
GW	.068804	.061809	.064468

Table 2. The true, normal and Saddlepoint p-values for the melanoma data

This data set is small but with heavy censoring. The saddlepoint approximations are accurate for both log-rank and generalized Wilcoxon type trend tests.

Carcinogenicity experiment

Thomas et al. (1977), presented the following carcinogenicity experiment results for the survival times of three groups in days until the tumor was observed.

group	$\{l_i\}$	$\{n_i\}$	Survival Time
1	2	10	$41^*, 41^*, 47, 47^*, 47^*, 58, 58, 58, 100^*, 117$
2	1.5	10	$43^*, 44^*, 45^*, 67, 68^*, 136, 136, 150, 150, 150$
3	0	9	$73^*, 74^*, 75^*, 76, 76, 76^*, 99, 166, 246^*$

Table 3 shows the comparison of the normal and the saddlepoint p-values with the true mid-p-values.

Test stat.	Normal p	Sadpt. mid-p	True mid-p
LR	.022226	.027959	.028730
GW	.034020	.032226	.03280

Table 3. The true, normal, and saddlepoint p-values for the carcinogenicity data.

Note that this data set has ties from within the same group or between censored and uncensored data, each of which makes no difference for both the normal and saddlepoint approximations. This happens because the total weight sums do not change by rearranging the tied values.

3.6 Simulation Study

Simulation studies are used to show the accuracy of the saddlepoint approximation over a wide range of data types, numbers of groups, sample sizes, degrees of censoring and error distributions. Three error distributions are used to simulate data: logistic, extreme value and Weibull distributions. Each distribution was considered using various numbers of groups and group sizes. In each consideration, 1000 data sets 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-value. The censored data were selected at random, indepen-

dently 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 $i\beta$ was added as a location shift parameter. The value of β was chosen so as to approximately achieve a p-value of 5%. After arranging the values from all groups, ranks were assigned and used to calculate the weights for the statistics and to calculate the p-values in the formulas of in §§2-3. The doses were chosen to be equally spaced $\{0, 1, 2, ...\}$.

For each of the 1000 data sets, the simulated mid-p-value was calculated by using 10^6 permutations of the test statistic computed by holding the censored positions fixed. The simulated mid-p-value is then the proportion of such generations exceeding the observed statistic plus half the proportion of those equal. These calculations were implemented for both log-rank and generalized Wilcoxon type trend tests. Tables 4-6 show the results for the three 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 data sets, "Sad. Prop." is the proportion of the 1000 data sets for which the saddlepoint mid-p-value was closer to the true mid-p-value than the normal p-value, "Abs. Err. Sad." is the average absolute error of the saddlepoint p-value from the true mid-p-value, "Rel. Abs. Err. Sad." 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.

Stat.	Mean	Sad. Prop.	Abs. Err. Sad.	Abs. Err. Normal	Rel. Abs. Err. Sad	Rel. Abs. Err. Nor.
3 groups, sizes= 7, 8, 9.				$\beta = 1$		30% censoring
LR	.043	.945	$.0^3364$	$.0^2486$	$.0^3148$	$.0^3619$
GW	.035	.931	$.0^3131$	$.0^2232$	$.0^4259$	$.0^3284$
3 groups, sizes= 5, 15, 25.				$\beta = 1$		15% censoring
LR	.031	.973	$.0^3184$	0.0^2457	0.0^4138	0.0^3790
GW	.020	.924	$.0^3120$	0.0^2181	0.0^4333	0.0^3743
3 groups, sizes= 25, 20, 30.				$\beta = .225$		30% censoring
LR	.076	.948	0.0^3207	$.0^2314$	$.0^4176$	.00137
GW	.080	.919	0.0^3169	$.0^2183$	$.0^4666$	$.0^3529$
5 groups, sizes= 7, 8, 7, 7, 8.				$\beta = .45$		15% censoring
LR	.056	.956	$.0^3261$	$.0^2511$	0^5153	$.0^4372$
GW	.045	.935	$.0^3125$	$.0^2231$	$.0^5108$	$.0^3113$
5 groups, sizes= 5, 15, 15, 5, 10.				$\beta = .5$		15% censoring
LR	.041	.959	0^3130	$.0^2373$	$.0^4387$	$.0^3447$
GW	.029	.933	$.0^4895$	$.0^2194$	$.0^4210$	$.0^3673$

Table 4. Performance under simulation from the logistic distribution.

Stat.	Mean	Sad. Prop.	Abs. Err. Sad.	Abs. Err. Normal	Rel. Abs. Err. Sad	Rel. Abs. Err. Nor.
3 groups, sizes= 9, 7, 8.				$\beta = .65$		5% censoring
LR	.029	.945	$.0^3307$	$.0^2365$	$.0^5646$	$.0^3164$
GW	.043	.921	$.0^3259$	$.0^2263$	$.0^5177$	$.0^4255$
3 groups, sizes= 15, 5, 25.				$\beta = .15$		15% censoring
LR	.052	.975	$.0^3320$	$.0^2380$	0.0^5795	0.0^3154
GW	.071	.906	$.0^3299$	$.0^2172$	0.0^4541	0.0^4121
5 groups, sizes= 7, 8, 7, 7, 8				$\beta = .2$		30% censoring
LR	.070	.962	$.0^3318$	$.0^2566$	$.0^6181$	$.0^3323$
GW	.091	.960	$.0^3194$	$.0^2373$	$.0^5347$	$.0^4595$
5 groups, sizes= 5, 15, 15, 5, 10.				$\beta = .2$		30% censoring
LR	.049	.962	$.0^3153$	$.0^2437$	0^5313	$.0^4419$
GW	.067	.965	$.0^3146$	$.0^2399$	$.0^6523$	$.0^4189$
5 groups, sizes= 15, 12, 15, 10, 13.				$\beta = .15$		30% censoring
LR	.055	.962	0^3157	$.0^2396$	$.0^4867$	$.0^3244$
GW	.078	.949	$.0^3161$	$.0^2299$	$.0^4164$	$.0^3106$

Table 5. Performance under simulation from the extreme value distribution.

Stat.	Mean	Sad. Prop.	Abs. Err. Sad.	Abs. Err. Normal	Rel. Abs. Err. Sad	Rel. Abs. Err. Nor.
3 groups, sizes= 9, 8, 7.				$\beta = .3$		15% censoring
LR	.049	.945	$.0^3402$	$.0^2495$	$.0^5876$	$.0^3236$
GW	.030	.906	$.0^3184$	$.0^2186$	$.0^5454$	$.0^4132$
3 groups, sizes= 25, 5, 15.				$\beta = .15$		5% censoring
LR	.085	.947	$.0^3326$	$.0^2338$	$.0^6597$	$.0^4252$
GW	.057	.916	$.0^3152$	$.0^2131$	$.0^5135$	$.0^5341$
3 groups, sizes= 30, 20, 25.				$\beta = .1$		15% censoring
LR	.101	.950	$.0^3233$	$.0^2339$	$.0^5276$	$.0^4531$
GW	.083	.909	$.0^3162$	$.0^2161$	$.0^5245$	$.0^4133$
5 groups, sizes= 7, 8, 7, 7, 8.				$\beta = .1$		5% censoring
LR	.092	.959	$.0^3368$	$.0^2648$	0^5487	$.0^4819$
GW	.073	.940	$.0^3187$	$.0^2290$	$.0^6695$	$.0^5938$
5 groups, sizes= 5, 15, 5, 15, 10.				$\beta = .1$		30% censoring
LR	.055	.973	$.0^3157$	$.0^2483$	$.0^4300$	$.0^3138$
GW	.042	.944	$.0^3116$	$.0^2246$	$.0^4817$	$.0^3191$

Table 6. Performance under simulation from the Weibull distribution.

3.6.1 Conclusion

For the log-rank simulations, the saddlepoint mid-p-value was more accurate in 95.7% 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 93.05% of the overall cases. In all three tables, the average absolute saddlepoint error was less than 10^{-3} with average relative error typically less than 0.01%.

Chapter 4
Generalization of Two Sample and Trend Tests to The Weighted Log-Rank Class

4.1 Introduction

Weighted log-rank statistics have been widely used in medical and epidemiological follow-up and survival analysis studies. The original (unweighted) log-rank test was proposed by Mantel and Haenszel (1959) and coincides with the partial likelihood score of Cox (1972) for proportional hazards regression model. On another front, efforts to extend the Wilcoxon rank sum test to censored failure time data led to the work of Gehan (1965), Peto and Peto (1972), Prentice (1978), the Tarone and Ware (1977) class and the Fleming and Harrington (1981) class.

Both the log-rank statistic and the extensions of the Wilcoxon statistic can be incorporated into the class of weighted log-rank statistics. This chapter generalizes the methods represented in chapters 2 and 3 for two sample and trend tests to the weighted log-rank class.

4.2 Weighted Log-rank Class

Let $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ represent the distinct ordered failure times in a sample of independent right censored data of $p+1$ groups pooled. Further let $t_{i1}, \cdots, t_{im_i}$ be the

right censored times in the intervals $[t_{(i)}, t_{(i+1)}),\ i = 0, 1, \cdots, k$ where $t_{(0)} = -\infty$, $t_{(k+1)} = \infty$ and $k + \sum_{i=1}^{k} m_i = n$. Also let $z_{(i)}$ and z_{ij} represent the corresponding indicator vector of group membership. Assuming no ties among the uncensored data from different groups, the general weighted log-rank class of statistics can be written as

$$v = \sum_{i=1}^{k} w_i \left(z_{(i)} - \frac{1}{n_i} \sum_{l \in R(t_{(i)})} z_l \right) \tag{4.1}$$

where n_i is the total number of individuals at risk at time $t_{(i)}^{-}$, and $R(t_{(i)})$ is the set of individuals at risk at that time. In most applications w_i is a fixed function of the risk set sizes $\{n_1, n_2, \cdots, n_i\}$ up to time $t_{(i)}$.

A variety of weight functions $\{w_i\}$ have been proposed in the literature. A common weight function, $w_i = 1$, leads to the log-rank test which has optimum power to detect alternatives where the hazard rates in the $p+1$ populations are proportional to one another. Another choice $w_i = n_i$, yields Gehan's (1965b) generalization of Mann-Whitney (1947) and Breslow's (1970) generalization of the Kruskal-Wallis (1952) test. Tarone and Ware (1977) suggest the weight function $w_i = f(n_i)$, with f as a fixed function, and further recommend the choice $f(y) = y^{1/2}$. As compares to log-rank this test gives more weight to differences between the observed and expected number of deaths at time points where there is the most data, but less weight than the Gehan test.

Peto and Peto (1972) and Kalbfleish and Prentice (2002) proposed the weight function that uses an estimate of the survival function,

$$w_i = \widetilde{S}(t_i) = \prod_{j=1}^{i} \frac{n_j}{n_j + 1}.$$

Andersen et. al. (1982) suggest that this weight should be modified slightly as

$$w_i = \widetilde{S}(t_i)\frac{n_i}{n_i + 1}$$

Fleming and Harrington (1981) proposed a general class of tests with weight function

$$w_i = \hat{S}(t_{i-1})^p \left\{1 - \hat{S}(t_{i-1})\right\}^q \qquad p \geqslant 0, q \geqslant 0.$$

here $\hat{S}(t_i)$ is the product limit estimator at time t_i,

$$\hat{S}(t) = \prod_{t_i \leq t}(1 - 1/n_i), \qquad t_1 \leq t.$$

The estimated variance-covariance matrix V_0, of vector v in (4.1), has components

$$\widehat{\sigma}_{jj} = \sum_{i=1}^{k} w_i^2 \frac{n_{ij}}{n_i}(1 - \frac{n_{ij}}{n_i})$$

and

$$\widehat{\sigma}_{jl} = -\sum_{i=1}^{k} w_i^2 \frac{n_{ij}}{n_i}\frac{n_{il}}{n_i}$$

for $j, l \in \{1, ..., p+1\}$, where n_{ij} is the number at risk from group j at time $t_{(i)}^-$. The normal approximation uses the asymptotic normality of $(v^-)^T(V_0^-)^{-1}v^-$, where v^- consists of the first p components of v and V_0^- is the upper left $(p \times p)$ block of V_0.

4.2.1 The Permutation Distribution of the Weighted Log-Rank Class

To generalize the methods represented in Chapters 2 and 3 to the weighted log-rank class, we need to rewrite v in the linear form

$$v = \sum_{i=1}^{k} \left\{ c_i z_{(i)} + C_i \sum_{j=1}^{m_i} z_{ij} \right\} \tag{4.2}$$

and build the saddlepoint approximation upon the permutation distribution of v as recorded below.

Proposition The weighted log-rank statistic v in (4.1) has its null distribution given as the distribution of $\sum_{i=1}^{k}(c_i z_{(i)} + C_i \sum_{j=1}^{m_i} z_{ij})$, where $z_1, ..., z_n$ are $(p+1)\times 1$ vectors that have a uniform distribution over all one way designs for $(p+1)\times 1$ vectors such that $\sum_{i=1}^{k} z_i = (n_1, ..., n_{p+1})^T$. This uniform distribution places probability $\binom{n}{n_1, ..., n_{p+1}}^{-1}$ on each design. The weights in (4.2) are

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

Proof.

$$\begin{aligned} v &= \sum_{i=1}^{k} w_i \Big(z_{(i)} - \frac{1}{n_i} \sum_{l \in R(t_{(i)})} z_l \Big) \\ &= \sum_{i=1}^{k} w_i z_{(i)} - \sum_{i=1}^{k} \frac{w_i}{n_i} \sum_{l \in R(t_{(i)})} z_l \end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{k}\frac{w_i}{n_i}\sum_{l\in R(t_{(i)})} z_l &= \frac{w_1}{n_1}\left(z_{(1)}+\sum_{j=1}^{m_1} z_{1j}+z_{(2)}+\sum_{j=1}^{m_2} z_{2j}+\ldots\right)\\ &\quad+\frac{w_2}{n_2}\left(z_{(2)}+\sum_{j=1}^{m_2} z_{2j}+z_{(3)}+\sum_{j=1}^{m_3} z_{3j}+\ldots\right)\\ &\quad+\frac{w_3}{n_3}\left(z_{(3)}+\sum_{j=1}^{m_3} z_{3j}+z_{(4)}+\sum_{j=1}^{m_4} z_{4j}+\ldots\right)\\ &= \sum_{i=1}^{k}\left(\sum_{l=1}^{i}\frac{w_l}{n_l}\right)\left\{z_{(i)}+\sum_{j=1}^{m_i} z_{ij}\right\}.\end{aligned}$$

Thus,

$$v=\sum_{i=1}^{k}\left\{\left(w_i-\sum_{l=1}^{i}\frac{w_l}{n_l}\right)z_{(i)}+\left(-\sum_{l=1}^{i}\frac{w_l}{n_l}\right)\sum_{j=1}^{m_i} z_{ij}\right\}.$$

■

The following table shows w_i and the corresponding c_i and C_i for some important test statistics.

Test	w_i	c_i	C_i
log-rank	1	$1-\sum_{j=1}^{i}\frac{1}{n_j}$	$-\sum_{j=1}^{i}\frac{1}{n_j}$
Gehan	n_i	n_i-i	$-i$
Peto-Prentice	$\prod_{j=1}^{i}\frac{n_j}{n_j+1}$	$2\prod_{j=1}^{i}\frac{n_j}{n_j+1}-1$	$\prod_{j=1}^{i}\frac{n_j}{n_j+1}-1$
Tarone-Ware	$\sqrt{n_i}$	$\sqrt{n_i}-\sum_{j=1}^{i}\frac{1}{\sqrt{n_j}}$	$-\sum_{j=1}^{i}\frac{1}{\sqrt{n_j}}$
Fleming-Harrington	$\hat{S}(t_{i-1})$	$\hat{S}(t_{i-1})-\sum_{j=1}^{i}\frac{\hat{S}(t_{j-1})}{n_j}$	$-\sum_{j=1}^{i}\frac{\hat{S}(t_{j-1})}{n_j}$

Table 1. The Weights w_i for the various tests along with the corresponding c_i and C_i.

Note that statistic (4.1) is the generalization to the censored case of test statistics that would be used in the uncensored setting. In the uncensored case, taking c_i as the expectation of the i^{th} order statistic of the normal distribution leads to the normal score test (Fisher and Yates, 1963). The van der Waerden (1953) test derives from

taking $c_i = \Phi\{i/(n+1)\}$. The choice c_i =sgn$\{2i-(n+1)\}$, $c_i = 0$ if $i = (n+1)/2$ leads to the Sign (median) score test.

Using the form (4.2) of the weighted log-rank class, we can summarize the saddlepoint approximation for the permutation distribution of v in the case of the two sample and trend tests as follows in the next two subsections.

4.2.2 Saddlepoint Approximation for the Two Sample Tests

For two sample tests, the $\{z_i\}$ in (4.2) are 1-dimension indicator vectors. As in Theorem 1 Chapter 2, v has the same distribution as $\sum_{i=1}^{k} q_i\xi_i | \sum_{i=1}^{k} \xi_i = n_1$ where $\{q_i\}$ are $\{c_i\}$ for uncensored times and $\{q_i\}$ are $\{C_i\}$ for censored times. Therefore the saddlepoint method can be applied in the following steps

1. Solve the saddlepoint equations

$$K'_s(\hat{s},\hat{t}) = \sum_{i=1}^{n} \frac{\exp(\hat{s}+q_i\hat{t})}{(1-\theta)/\theta + \exp(\hat{s}+q_i\hat{t})} = n_1 \tag{4.3}$$

$$K'_t(\hat{s},\hat{t}) = \sum_{i=1}^{n} \frac{q_i\exp(\hat{s}+q_i\hat{t})}{(1-\theta)/\theta + \exp(\hat{s}+q_i\hat{t})} = v_0 \tag{4.4}$$

where n_1 is the number of individuals in the treatment group , $\theta = n_1/n$, v_0 is the observed test statistic and $s_0 = 0$. Solution of (4.3) and (4.4) provides the saddlepoint $(\hat{s},\hat{t})$.

2. The saddlepoint equations use the CGF

$$K(s,t) = \sum_{i=1}^{n} \log\left\{1 - \frac{n_1}{n} + \frac{n_1}{n}\exp(s+q_it)\right\}.$$

The inputs to the Skovgaard approximation are

$$\hat{w} = \text{sgn}(\hat{t})\sqrt{2\left[\{K\left(\hat{s}_0, 0\right) - n_1\hat{s}_0\} - \{K\left(\hat{s}, \hat{t}\right) - n_1\hat{s} - v_0\hat{t}\}\right]}$$

$$\hat{u} = \hat{t}\sqrt{\left|K''(\hat{s}, \hat{t})\right| / K''_{ss}(\hat{s}_0, 0)}.$$

In these expressions, K'' is the 2×2 Hessian matrix and K''_{ss} is the $\partial^2/\partial s^2$ component of this Hessian.

3. The Skovgaard approximation for testing $H_0 : \beta = 0$ versus $H_1 : \beta > 0$ has one-sided mid-p-value that is approximated as

$$\Pr(v \geq v_0 | \sum_{i=1}^{k} z_i = n_1) \simeq 1 - \Phi(\hat{w}) - \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right).$$

Example Consider the Pike (1966) vaginal cancer data, in Table 5 of Chapter 2. We compare the p-values for saddlepoint and normal approximations with the true mid-p-values of the Gehan, Tarone-Ware, and Fleming and Harrington tests. The calculations for log-rank and Peto-Prentice generalization for Wilcoxon test have been provided in Chapter 2.

We should note here that the weights given for log-rank and generalized Wilcoxon in Table 1 are the negative values of the weights given in Chapter 2. This sign does not affect the normal calculation because of the symmetry. It also does not affect the saddlepoint calculations. When we change the sign of $\{q_i\}$ in the saddlepoint equations and the value of the statistic to its negative value, the resulting saddlepoint is $\hat{s}$ and $-\hat{t}$, but the values of $\hat{w}$ and $\hat{u}$ for this saddlepoint are the same as for $\hat{s}$ and $\hat{t}$, which results in the same mid-p-value. Table 2 shows the comparison of

the saddlepoint and the normal approximations based on the true (simulated) mid-p-value for the Gehan, Tarone-Ware, and Fleming and Harrington two sample tests.

Test	True mid-p	Sadpt. mid-p	Normal p
Gehan	.060366	.060435	.068348
Tarone-Ware	.054247	.054118	.064380
Fleming-Harrington	.052096	.052425	.058946

Table 2. Comparison of saddlepoint and normal p-values for the Pike data.

Table 2 represent an example of the accuracy achieved by the saddlepoint method. The saddlepoint approximation appears to achieve the same accuracy as seen with the log-rank and generalized Wilcoxon tests proposed in Chapters 2.

4.2.3 Saddlepoint Approximation for Tests for Trend

The general form of the trend tests based on the weighted log-rank tests can be specified as

$$u = l^T v = l^T \sum_{i=1}^{k} (c_i z_{(i)} + C_i \sum_{j=1}^{m_i} z_{ij})$$

where $l^T = (l_1, ..., l_{p+1})$ is vector of doses and $\{z_i\}$ are $(p+1)\times 1$ vector indicators of group membership. Using Proposition 1 and Theorem 1 in Chapter 3, the saddlepoint approximation for the mid-p-value in testing $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p+1}$ vs $H_1 : \beta_1 \leq \beta_2 \leq \cdots \leq \beta_{p+1}$ may be summarized in the following steps.

1. Solve the saddlepoint equations,

$$K'_{s_l}(\hat{s}, \hat{t}) = \sum_{i=1}^{n} \frac{n_l \exp(\hat{s}_l + r_{il}\hat{t})}{\left\{\sum_{j=1}^{p} n_j \exp(s_j + r_{ij}t) + n_{p+1}\right\}} = n_l, \qquad l = 1, ..., p$$

$$K_t'(\hat{s},\hat{t}) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{p} n_j r_{ij} \exp(s_j + r_{ij}t)}{\left\{\sum_{j=1}^{p} n_j \exp(s_j + r_{ij}t) + n_{p+1}\right\}} = u_0$$

where $r_{ij} = q_i(l_j - l_{k+1})$, $\{q_i\}$ are $\{c_i\}$ for the uncensored times, $\{q_i\}$ are $\{C_i\}$ for censored times, and u_0 is the observed statistic. The solution leads to $\hat{s}^T = (\hat{s}_1, ..., \hat{s}_p)$, $\hat{t}$ and $\hat{s}_0 = (0, ..., 0)$.

2. The saddlepoint equations use $(p+1)$-dimensional CGF

$$K(s,t) = \sum_{i=1}^{n} \log\left[\left\{\sum_{j=1}^{p} \frac{n_j}{n} \exp(r_{ij}t + s_j)\right\} + \frac{n_{p+1}}{n}\right]$$

where n_j is the number of individual in group j and $N = (n_1, ..., n_p)^T$. The inputs to the Skovgaard approximation are

$$\hat{w} = \text{sgn}(\hat{t})\sqrt{2\left[\{K(\hat{s}_0, 0) - \hat{s}_0^T N\} - \{K(\hat{s},\hat{t}) - \hat{s}^T N - u_0\hat{t}\}\right]}$$

$$\hat{u} = \hat{t}\sqrt{|K''(\hat{s},\hat{t})|/|K''_{ss}(\hat{s}_0, 0)|}.$$

In these expressions, K'' is the $(p+1) \times (p+1)$ Hessian matrix and K''_{ss} is the $\partial^2/\partial s \partial s^T$ portion at $(\hat{s}_0, 0)$.

3. The Skovgaard approximation to the mid-p-value is

$$\Pr(u \geq u_0) = 1 - \Phi(\hat{w}) - \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right)$$

Example The data are from the carcinogenicity experiment of Thomas et. al. (1977) given in Table 3 of Chapter 3. Table 3 summarizes the mid-p-values for the true, saddlepoint approximation and the normal for Gehan, Tarone-Ware, and Fleming-Harrington tests for the one sided test for trend of three group.

Test	True mid-p	Sadpt. mid-p	Normal p
Gehan	.018591	.018222	.025377
Tarone-Ware	.023662	.023613	.025578
Fleming-Harrington	.033245	.033237	.035962

Table 3. Saddlepoint and normal mid-p-values approximations for the carcinogenicity data.

Table 3 shows the high accuracy of the saddlepoint approximation as compares to the normal approximation in the three trend tests. The accuracy shows 3 or 4 digits.

4.3 Tied Data in Weighted Log-Rank Class

Suppose $t_{(1)} < ... < t_{(k)}$ are the ordered event times, and at each event time $t_{(i)}$ there are d_i individuals that died with z_{ij} as the is indicator vector of group membership (in $p+1$ groups) for individual j at time i, $j = 1, ..., d_i$ and $i = 1, ..., k$. Let $z'_{i1}, ..., z'_{im_i}$ be indicator vectors of group membership for the censored data in $[t_{(i)}, t_{(i+1)})$ with a total of n individuals in the trial.

The weighted log-rank statistic for the tied data is

$$v_t = \sum_{i=1}^{k} w_i \left(\sum_{j=1}^{d_i} z_{ij} - \frac{d_i}{n_i} \sum_{l \in R(t_{(i)})} z_l \right) \tag{4.5}$$

where n_i is the number of individuals at risk at time $t^-_{(i)}$, $R(t_{(i)})$ is the set of individuals at risk at time $t^-_{(i)}$, see Kabfliesch and Prentice (2002, eq.4.25) and Klein and Moeschberger (1997, eq. 7.3.3). It is easy to show that the statistic v_t (as in the proof of the proposition) can be written in the form of (4.2) as

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

with

$$c_i = \left(w_i - \sum_{j=1}^{i} w_j d_j / n_j \right)$$

and

$$C_i = \left(- \sum_{j=1}^{i} w_j d_j / n_j \right).$$

From this form, one can see that v_t is still linear in z_{ij} and z'_{il}. Also the tied values at $t_{(i)}$ have the same weight, which means that this general setting for the weighted log-rank class in the case of ties is equivalent to the score average setting explained in Chapter 2. Since the ties do not affect the linearity of v_t, the permutation distribution of v_t can be approximated by the conditional Skovgaard method as

$$\Pr(v_t > v_0) = \Pr(\sum_{i=1}^{n} q_i \zeta_i | \sum_{i=1}^{n} \zeta_i = N).$$

Here $\zeta_1, ..., \zeta_n$ are Bernoulli (θ) and $N = n_1$ is the number of points in the treatment group, for the two sample problem. For the trend test, $\zeta_1, ..., \zeta_n$ are Multinomial $(1; \theta_1, ..., \theta_p)$ with $N = (n_1, ..., n_{p+1})^T$ as a vector denoting the number of points in each group. The saddlepoint approximations for both cases are the same as explained in the previous section except that the weights are changed to

$$c_i = \left(w_i - \sum_{j=1}^{i} w_j \frac{d_j}{n_j} \right)$$

for the death points and to

$$C_i = \left(- \sum_{j=1}^{i} w_j \frac{d_j}{n_j} \right)$$

for the censored points.

Two Sample Example; This example compares the saddlepoint and the normal approximations with the true mid-p-value for the data recording time to kidney infection of Nahman et. al. (1992). Five tests are used including log-rank, Gehan, Peto-Prentice generalized Wilcoxon, Tarone-Ware and Fleming and Harrington tests. The data consists of 119 individuals of which 43 have surgically placed catheter and 76 have a percutaneous placed catheter. The saddlepoint approximations are extremely accurate.

mid-p-value	L-R	Ghn	G.W.	T-W	H-F
True	.050982	.488313	.113630	.257416	.114372
Sadpt.	.051222	.489087	.113398	.256913	.114381
Normal	.055867	.481792	.118432	.262839	.119496

Examples of Tests for Trend Approximations for mid-p-values are presented when considering tests for trend using the same five tests. Two data sets are considered. The first set of data are from the carcinogenicity experiment of Thomas et. al. (1977). This data consists of three groups with a total of 29 patients in all. A second data set, given in Schmee and Hahn (1979), is a study of the accelerated life tests on electrical insulation, which has four groups with ten motorettes tested.

mid-p-value	L-R	Ghn	G.W.	T-W	H-F
The carcinogenicity experiment					
True	.021511	.007877	.016681	.012518	.015034
Sadpt.	.021439	.007793	.016492	.012525	.014981
Normal	.009780	.007070	.011765	.008326	.010536
Electrical isolation data					
True	$.0^468$	$.0^420$	$.0^442$	$.0^420$	$.0^436$
Sadpt.	$.0^467$	$.0^420$	$.0^434$	$.0^428$	$.0^432$
Normal	$.0^683$	$.0^412$	$.0^533$	$.0^535$	$.0^542$

The two tables show high accuracy for the saddlepoint approximations in both the two sample tests and the trend tests. The three data sets represent both light and heavy censoring, small and large number of failure times, and various numbers of groups all with ties. The saddlepoint approximation is always more accurate than the normal approximation. Generally we can apply the tests in this setting for the ties or we can use the permutation method suggested in Chapter 2.

Chapter 5
Tests for Independence and Symmetry

5.1 Testing For Independence

5.1.1 Linear Rank Tests for Independence

Some of the most important applications of the trend tests occur in situations when the factors being studied are not treatments that the investigator can assign to his subjects but conditions or attributes which are inseparably attached to these subjects. For example, it is not possible to assign people at an early age to various smoking habits in order to study the effect of smoking on a person's health. The hypothesis to be tested is that an association exists between two factors in a population of subjects, which may be people, manufactured items, institutions, and so on. With each subject are associated two characteristics: a person's ability for mathematics and music, the size and crime rate of a city and so forth. To test the hypothesis of independence that there is no relationship between the two characteristics, a sample of N subjects is drawn from the population and the values of the two characteristics are obtained for each member of the sample. Lehmann (1975) illustrate how these data can be used to test the hypothesis of independence using the following example.

Example From a group of 98 students enrolled in a statistics course, nine are selected at random and given a simple arithmetic and language tests with the following results

code of student	74	91	33	27	76	29	09	25	67
language	50	23	28	34	14	54	46	52	53
arithmetic	38	28	14	26	18	40	23	30	27

Let us rank separately the scores on each of the two tests ,

code of student	74	91	33	27	76	29	09	25	67
language	6	2	3	4	1	9	5	7	8
arithmetic	8	6	1	4	2	9	3	7	5

A clearer view of the relationship of the two sets of ranks is obtained by arranging one of them, say the first, in its natural order,

language	1	2	3	4	5	6	7	8	9
arithmetic	2	6	1	4	3	8	7	5	9

We denote the ranks in the second row by $(T_1, ..., T_N)$. Under the assumption of independence, all $N!$ orderings $(T_1, ..., T_N)$ are equally likely with probability $1/N!$. A formal proof of this fact is given at Chapter 7 in Lehmann (1975). If we are willing to assume that the two scores have a positive association, the $\{T_i\}$ should reveal an upward trend, with large values tending to occur on the right of the second row and low values on the left. The test statistic

$$D = \sum_{i=1}^{N} (T_i - i)^2 \tag{5.1}$$

would be appropriate, with small values of D indicating significance. Under the assumption of independence,

$$E(D) = \frac{N^3 - N}{6},$$

and

$$Var(D) = \frac{N^2(N+1)^2(N-1)}{36}.$$

The statistic D is related to Spearman's coefficient of rank correlation, R, see Gibbons and Chakraborti (2003), with the relation

$$R = 1 - \frac{6D}{N(N^2-1)}.$$

It is also related to the weighted Mann statistic D', by

$$D' = \frac{1}{6}N(N^2-1) - \frac{1}{2}D$$

where

$$D' = \sum_{i<j}(j-i)U_{ij}$$

and $U_{ij} = 1$ or 0 as $T_i < T_j$ or $T_i > T_j$.

By expanding (5.1), D can be written as

$$D = \frac{1}{3}N(N+1)(2N+1) - 2\sum_{i=1}^{N} iT_i$$

which gives an equivalent simple statistic,

$$v' = \sum_{i=1}^{N} iT_i, \tag{5.2}$$

see Hajek, Sidak and Sen (1999). Under independence,

$$Ev' = \frac{1}{4}N(N+1)^2,$$

and

$$Var(v') = \frac{N^2(N+1)^2(N-1)}{144}.$$

Taking $T = (T_1, ..., T_N)^T$, the statistic (5.2) can be rewritten as

$$\begin{aligned} v' &= l^T T, \qquad l^T = (1, 2, ..., N) \\ &= l^T \sum_{i=1}^{N} i z_i \end{aligned}$$

where $z_1, ..., z_n$ are $N \times 1$ vectors of the following form. Let the $(N \times N)$ identity matrix $I_N = (\eta_1, ..., \eta_N)$. Then

$$z_{T_i} = \eta_i, \qquad i = 1, ..., N$$

For example, $T_1 = 2$ in arithmetical rank so that $z_2 = \eta_1$ and $\sum_{i=1}^{N} i z_i$ has a 2 in its first component for T_1. This can be generalized to the form

$$v = l^T \sum_{i=1}^{N} q_i z_i \tag{5.3}$$

where q_i's are constant functions of $1, ..., N$, which is the usual form of the simple linear rank score tests with no ties. Cox and Oakes (1984) mentioned that this type of test can be modified for censored data, rather as Gehan modified the Wilcoxon-Mann-Whitney test. That means in the censored case the statistic (5.3) has the same form but with $\{q_i\}$ as a function of the number at risk up to time $t_{(i)}$, and still represented as a linear function in the $\{z_i\}$. Shirahata (1975) derived a locally most powerful test for independence with censored data when the smallest n_1 observations of the first characteristic and the smallest n_2 observations from second characteristic are uncensored and the rest of the data are censored. He used a score test statistic and

showed the asymptotic normality of the score test. Another form of Fisher-Yates (normal score) test has been given by Vorlickova (1976) when ties existence. Both tests do not fit into this framework. Also it seems hard to implement the Cox and Oakes idea in this linear form.

5.1.2 Saddlepoint Approximation for Tests of Independence

To deal with the statistic v in (5.3) using the saddlepoint approximation, note that

$$z_1, ..., z_N \stackrel{D}{=} \zeta_1, ..., \zeta_N | \sum_{i=1}^{N} \zeta_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

where $\zeta_1, ..., \zeta_N$ are $N \times 1$ i.i.d. vectors of Multinomial $(1, \theta_1, ..., \theta_N)$ which is a special case of the derivations of Chapter 3 when $n_1 = n_2 = ... = n_{p+1} = 1$ and $p+1 = N$. The dependence in the statistic can be removed by using the $(N-1) \times 1$ vectors z_i^- and ζ_i^-, the first $N-1$ components in z_i and ζ_i, so

$$z_1^-, ..., z_n^- \stackrel{D}{=} \zeta_1^-, ..., \zeta_n^- | \sum_{i=1}^{n} \zeta_i^- = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

v can be written as a function of $z_1^-, ..., z_n^-$ as

$$v = l_-^T \sum_{i=1}^{N} q_i z_i^- + Q$$

where $l_-^T = (l_1 - l_N, ..., l_{N-1} - l_N)$ and $Q = l_N \sum_{i=1}^{N} q_i$. Then by theorem 2 in Chapter 3, the null distribution of v is

$$\Pr\{v \geq v_0\} = \Pr\{l_-^T \sum_{i=1}^{N} q_i \zeta_i^- + Q \geq v_0 \,|\, \sum_{i=1}^{N} \zeta_i^- = (1, ..., 1)^T\}$$

which can be approximated using Skovgaard saddlepoint approximation by solving the saddlepoint equations

$$K'_{s_l}(\hat{s},\hat{t}) = \sum_{i=1}^{N} \frac{\exp(\hat{s}_l + r_{il}\hat{t})}{\left\{\sum_{j=1}^{N-1}\exp\left(s_j + r_{ij}t\right)+1\right\}} = 1, \qquad l = 1,...,N-1$$

$$K'_t(\hat{s},\hat{t}) = \sum_{i=1}^{N} \frac{\sum_{j=1}^{N-1} r_{ij}\exp\left(s_j + r_{ij}t\right)}{\left\{\sum_{j=1}^{N-1}\exp\left(s_j + r_{ij}t\right)+1\right\}} = v_0$$

in the $(N-1)$ components of $\hat{s}$ and scalar $\hat{t}$. The mid-p-value for the independence test is

$$\Pr(v \geq v_0) \simeq 1 - \Phi(\hat{w}) - \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right)$$

where

$$\hat{w} = \operatorname{sgn}(\hat{t})\sqrt{2\left[-\{K\left(\hat{s},\hat{t}\right) - \hat{s}^T x - v_0\hat{t}\}\right]}$$

$$\hat{u} = \hat{t}\sqrt{|K''(\hat{s},\hat{t})|/|K''_{ss}(0,0)|}.$$

and $x = (1,...,1)^T$. In these expressions, K'' is the $N \times N$ Hessian matrix and K''_{ss} is the $\partial^2/\partial s \partial s^T$ portion at $(0,0)$. An important consideration in these saddlepoint computations is the difficulty in solving N saddlepoint equations. This becomes increasingly difficult with large N.

Example Nayak (1988), gives the failure times of transmission (X) and of transmission pumps (Y) on 15 caterpillar tractors. To test the independence of failure times of X and Y, the test statistic (5.2)

X	1641	5556	5421	3168	1534	6367	9460	6679
	6142	5995	3953	6922	4210	5161	4732	
Y	850	1607	2225	3223	3379	3832	3871	4142
	4300	4789	6310	6311	6378	6449	6949	

can be used with $l = (1, ..., N)$ and $q_i = i$ and $Q = l_N \sum_{i=1}^{N} q_i = N^2(N+1)/2$. The true (simulated) mid-p-value $= 0.276807$ and the saddlepoint approximated mid-p-value $= 0.276303$. The normal mid-p-value is 0.269376.

5.2 Tests for Symmetry

5.2.1 Linear Rank Tests for Symmetry

Let $x_1, ..., x_N$ be i.i.d. random variables with distribution function F and density f. Classical statistical methods often assume symmetry of f about some value M, that is $f(M - x) = f(M + x)$ for all x. Thus it may be important to test such an assumption. Some authors consider the value M as the known median, $M = 0$; see Hajek, Sidak and Sen (1999), I.H. Tajuddin (1994) and Fellingham and Stoker (1964). Others used an estimated median such as the empirical median and applying the symmetry test about zero after centering the data using the estimated median; see Antille et. al. (1982).

Suppose $x_1, x_2, ..., x_N$ are the centered data set. The distribution F is assumed continuous so no observation assumes the exact value zero and no two observations have equal magnitude. Suppose we are interested in testing $H_0 : F(x) = 1 - F(-x)$ for all x vs $H_1 : F(x) \neq 1 - F(-x)$, for some x. Let R_i^+ be the rank of $|x_i|$ when the sequence $|x_1|, ..., |x_N|$ is arranged in ascending order. Hajek, Sidak and Sen (1999) represent some locally most powerful rank tests such as Fraser (normal

score) test with the statistic

$$S^+ = \sum_{X_i>0} a_N^+(R_i^+)$$

where $a_N^+(i) = E\Phi^{-1}(.5+.5U_N^{(i)})$, with Φ being the standardized normal distribution function and $U_N^{(1)}, ..., U_N^{(N)}$ being the ordered sample from the uniform distribution on $[0, 1]$. The Van der Waerden statistic is

$$S^+ = \sum_{X_i>0} \Phi^{-1}(.5 + .5\frac{R_i^+}{N+1})$$

and the Wilcoxon one sample test has statistic

$$S^+ = \sum_{X_i>0} R_i^+ \tag{5.4}$$

with mean and variance

$$\begin{aligned} E\left(S^+\right) &= \frac{1}{4}N(N+1) \\ Var(S^+) &= \frac{1}{24}N(N+1)(2N+1) \qquad \text{under } H_0. \end{aligned}$$

An equivalent test statistic is the Wilcoxon matched pair signed rank test is

$$T = 2\sum_{X_i>0} R_i^+ - \frac{1}{2}N(N+1)$$

with zero mean and variance

$$Var(T) = \frac{1}{6}N(N+1)(2N+1)$$

under H_0; see Van Eeden and Benard (1957). The Wilcoxon test statistic S^+ can be written as

$$S^+ = \sum_{i=1}^{N} R_i^+ z_i$$

with z_i as an indicator of $x_i > 0$; also the statistic T can be written as $T = \sum_{i=1}^{N} R_i^+(2z_i - 1)$. Another class of tests proposed by Antille et. al. (1982) has form

$$T_1(\alpha) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} G_\alpha \left\{ \frac{R_i^+}{2(N+1)} \right\} sgn(x_i)$$

where $G_\alpha(x) = \min(x, .5 - \alpha)$, with $0 \leq x \leq .5$ and $0 \leq \alpha \leq .5$; for $\alpha = 0$ this is Gupta's statistic. Also $T_1(\alpha)$ can be written as

$$T_1(\alpha) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} G_\alpha(\frac{R_i^+}{2(N+1)})(2z_i - 1)$$

The general linear rank tests representation of Hajek and Sidak (1967) is

$$S^+ = \sum_{X_i > 0} a_N^+(R_i^+)$$

and can be rewritten as $S^+ = \sum a_N^+(R_i^+) z_i$. The signed version of the statistic is $S^+ = \sum a_N^+(R_i^+)(2z_i - 1)$.

From the above discussion we can see that a large class of symmetry tests can be represented as

$$T^+ = \sum_{i=1}^{N} q_i z_i + C \tag{5.5}$$

where $\{q_i\}$ are the scores. For the Wilcoxon one sample test $q_i = R_i^+$; C is zero for the unsigned tests (Wilcoxon one sample test (5.4)) and $0 \neq C = \sum_{i=1}^{N} R_i^+ = N(N+1)/2$ for the Wilcoxon signed test.

5.2.2 Saddlepoint Approximation for the Symmetry Tests

The statistic in (5.5) is the same as the two sample test statistic in Chapter 2, the only difference is that z_i is the indicator that $x_i > 0$ rather than the indicator of the

treatment group membership. Based on this argument, the saddlepoint approximation for the two sample tests of Chapter 2 can be used to approximate the one-sided mid-p-value for testing symmetry. Here the weights $\{q_i\}$ specify the particular test statistic and C makes a slightly change in the saddlepoint equations.

The saddlepoint approximation for the mid-p-value of the statistic (5.5) can be calculated by solving the saddlepoint equations

$$K'_s(\hat{s},\hat{t}) = \sum_{i=1}^{n} \frac{\exp(\hat{s}+q_i\hat{t})}{n_2+n_1\exp(\hat{s}+q_i\hat{t})} = 1$$
$$K'_t(\hat{s},\hat{t}) = \sum_{i=1}^{n} \frac{q_i\exp(\hat{s}+q_i\hat{t})}{n_2+n_1\exp(\hat{s}+q_i\hat{t})} + C = t_0^+$$

to get $\hat{s}$ and $\hat{t}$, where t_0^+ is the observed statistic and n_1 is the number of x_i's greater than zero and $n_2 = n - n_1$. Then the mid-p-value for the symmetry test is approximated by the Skovgaard saddlepoint approximation

$$\Pr(T^+ \geq t_0^+) = 1 - \Phi(\hat{w}) - \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right)$$

where

$$\hat{w} = \operatorname{sgn}(\hat{t})\sqrt{2\left[\{K(\hat{s}_0,0) - n_1\hat{s}_0\} - \{K(\hat{s},\hat{t}) - n_1\hat{s} - t_0^+\hat{t}\}\right]}$$
$$\hat{u} = \hat{t}\sqrt{\left|K''(\hat{s},\hat{t})\right| / K''_{ss}(\hat{s}_0,0)}.$$

The group of tests presented in §5.2.1 are for uncensored data. The modifications and the generalizations of the Wilcoxon, log-rank and other classes of tests for two sample problem can be used with censored data to test for symmetry as Tajuddin (1994) used the Wilcoxon signed two sample test to test for symmetry. This gener-

alization has no effect on the permutation distribution of the statistic (5.5) and then the saddlepoint method still works. The only change would be in the weights corresponding to the censored and uncensored data, for example using the weights of Gehan statistic assign weights $n_i - i$ for uncensored data and $-i$ for the censored data, where n_i is the number of points have $|x|$ greater than $|x_i|$.

The idea of using the two sample tests for testing symmetry is to use $|x_1|, ..., |x_N|$ as the two sample data with treatment group consisting of the individuals whose $x_i > 0$. By this way Wilcoxon two sample test becomes Wilcoxon one sample test (5.4). This modification allow the use of all the two sample tests for censoring of Chapter 4 as tests for symmerty censored data. Two examples, with uncensored data and censored data, show the accuracy of saddlepoint approximation for symmetry.

Example (uncensored data) The data from Good and Gaskins (1980) are the percentage of silica calculated for 22 chondrites meteors and centered about 19.

−8.25	−6.44	−6.29	−6.01	−2.61	−1.92	−1.68	−1.67
−1.43	−1.19	−.31	.36	1.23	2.89	3.88	4.23
4.28	4.40	4.52	4.83	4.95	5.82		

The Wilcoxon signed rank test

$$T^+ = \sum_{i=1}^{N} 2R_i^+ z_i - \frac{1}{2}N(N+1)$$

is used to test symmetry with $q_i = 2R_i^+$ and $C = -\frac{1}{2}N(N+1)$. The statistic T^+ has mean zero and $var(T^+) = N(N+1)(2N+1)/6$. The true mid-$p$-value is .3615625. The saddlepoint mid-p-value is .3616631 and the normal mid-p-value is .42914118, the saddlepoint approximation is considerably more accurate.

Example (censored data) The following data are the asymptomatic part of data presented by Dinse, G. F. (1982) for survival times of patients with lymphocytic non-hodgkins lymphoma after centering around the median.

-247	-239	-201	-158	-145	-138	-108	-72	-58	-55
-40	-35	-16	-5	-3	3^*	4	9^*	32^*	45^*
49^*	52^*	57^*	62	63^*	65^*	68^*	81^*	84^*	91^*

To test the symmetry of this data we use as examples log-rank and Gehan tests. The following table shows true, saddlepoint, and normal mid-p-values for both tests. The high accuracy of saddlepoint approximation is made clear.

	True	Sad.pt.	normal
log-rank	.09225	.09033	.05051
Gehan	.05206	.05160	.04942

References

[1] Agresti, A. (1992). A survey of exact inference for contingency tables. *Statist. Sci.*, **7**, 131-153.

[2] Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1982). Linear nonparametric tests for comparison of counting processes, with application to censored survival data (with discussion). *Int. Stat. Rev.* ,**50**, 219-258.

[3] Antille, A., Kersting, G. and Zucchini, W. (1982). Testing symmetry. *JASA*, **77**, No. 379, 639-646.

[4] Barndorff-Nielsen, O. E. and Cox D. R. (1979). Edgeworth and saddlepoint approximations with statistical applications. *J. R. Statist. Soc.* **B 41**, 279-312.

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

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

[7] Butler, R.W. (2005). *Saddlepoint Approximations and Applications.*

[8] Chapman, P.L., Butler, R.W., and Paige, R.L. (2005). Saddlepoint confidence intervals for LD-100α. to be submitted to *J. Amer. Statist. Assoc.*

[9] Collett, D. (2003). *Modelling survival data in medical research.* 2ed. Chapman & Hall. New York.

[10] Cox, D.R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser.* **B, 34**, 187-220.

[11] Cox, D. R. and Oakes, D. (1984). *Analysis of survival data.* Chapman and Hall. New york.

[12] Cramer, H. (1938). *Actualites Scientifiques et Industrielles*, Vol. 736. Hermann, Paris.

[13] Cuzick, J. (1985), A wilcoxon-type test for trend, *Statistics in Medicine*, **4**, 87-90.

[14] Daniels, H. E. (1954). Saddlepoint approximation in statistics. *Ann. Math. Statist.*, **25**, 631-650.

[15] Daniels, H. E. (1956). The approximate distribution of serial correlation coefficients. *Biometrika*, **43**, 169-185.

[16] Daniels, H. E. (1958). Discussion of paper by D. R. Cox, *J. R. Statist. Soc.* **B 20**, 236-238.

[17] Daniels, H. E. (1987). Tail probability approximations. *Int. Statist. Rev.*, **55**, 37-48.

[18] Davison, A. C. and Hinkely D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika*, **75**,3 417-431.

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

[20] De Bruijn, N. G. (1980). *Asymptotic methods in analysis*. Dover, New York.

[21] Edmunson, J.H., Fleming,T.R., Decher, D.G., Malkasian, G.D., Jorgenson, E.O., Jeffries, J.A., Webb,M.J. and Kvols, L.K. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal resdual disease. *Cancer Treatment Reports* , **63**, 241-7.

[22] Fellingham, S. A. and Stoker, D. J. (1964). An approximation for the exact distribution of the Wilcoxon test for symmetry. *JASA*, **59**, 307, 899-905.

[23] Fisher, R. A. and Yates, F. (1963). *Statistical tables for biological, agricultural and medical research.* 6th edition. Edinburgh: Oliver and Boyed.

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

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

[26] Gehan, E. A. (1965b). A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika*, **52,** 650-652.

[27] Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric statistical inference*. 4th edition, Marcel Dekker, New York.

[28] Good, I. J. (1957). Saddlepoint methods for the multinomial distribution. *Ann. Math. Statist.*, **28**, 861-880.

[29] Good, I. J. (1961). The multivariate saddlepoint method and chi-squared for the multinomial distribution. *Ann. Math. Statist.*, **32**, 535-548.

[30] Good, I. S. and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method explained by scattering and meteorite data. *Journal of the American Stat.*, **A75**, 42-56.

[31] Hajek, J., Sidak, Z. and Sen, P. K. (1999). Theory of rank tests. 2nd Ed. Academic Press.

[32] Jonckheere A. R. (1954). A distribution-free K-sample test against ordered alternatives. *Biometrika*, **41**, 1, 133-145.

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

[34] Kim, D. and Agresti, A. (1995). Improved exact inference about conditional association in three-way contingency tables. *J. Amer. Statist. Assoc.*, **90**, 632-639.

[35] Kirk, A.P., Jain, S., Pocock, S. et al. (1980). Late results of the royal free hepatitis prospective controled trial of prednisolone therapy in hepatitis B surface antigen negative chronic active hepatitis. *Gut*, **21**,78-83.

[36] Klein, J.P., Moeschberger, M.L. (1997). *Survival analysis techniques for censored and truncated data.* New York, Springer, 10.

[37] Kruskal,W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion analysis of variance. *JASA*, **47**,583-721.

[38] Lee, E.T.(1992) *Statistical Methods for Survival Data Analysis*, Wiley, New York.

[39] Lehmann, E. L. (1975). *Nonparametrics, statistical methods based on ranks.* Holden-Day. San Francisco.

[40] Lugannani, R. and Rice, S. O. (1980). Saddlepoint approximations for the distribution of the sum of independent random variables. *Adv. Appl. Prob.*, **12**, 475-490.

[41] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, **18**, No. 1, 50-60.

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

[43] Mantel, N. and Haenszel, W. (1959), Statistical aspects of the analysis of data from restrospective studies of disease. *J. Nat. Cancer Inst.*, **22**, 719-748.

[44] Mehrotra, K. G., Michalek, J. E., and Mihalko, D. (1982). A relationship between two forms of linear rank procedures for censored data. *Biometrika*, **69**, 674-676.

[45] Nayak, T. K. (1988). Testing equality of conditionally independant exponential distributions. *Communications in statistics (theory and methods)*, **17,** 807-820.

[46] Nahman, N. S., Middendorf, D. F., Bay, W. H., McElligott, R., Powell, S., and Anderson, J.(1992). Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: clinical results in 78 patients. *Journal of the American Society of Nephrology*, **3**, 103-107.

[47] Page, E. B. (1963), Ordered hypotheses for multiple treatments: A significance test for linear ranks, *JASA*, **58**, 301, 216-230.

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

[49] Peto, R., Pike, M.C., Armitage, P., Breslow, N. E., Cox, D. R.,Howard, S.V., Mantal, N., McPherson, K., Peto, J., Smith, P.G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *British Journal of Cancer*, **35**, 1-39.

[50] Pierce, D.A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *J. R. Statist. Soc* **B, 54**, 701-737.

[51] Pike, M. C. (1966). A method of analysis of certain class of experiments in carcinogenesis. *Biometrics*, **22**, 142-161.

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

[53] Robinson, J. (1982). Saddlepoint approximations for permutation tests and confidence intervals. *J. R. Statist. Soc.* **B, 44**, No. 1, 91-101.

[54] Routledge, R.D. (1994). Practicing safe statistics with the mid-*p*. *Canadian J. Statist.*, **22**, 103-110.

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

[56] Sedmak, D.D., Meineke, T.A., Knechtges D.S., and Anderson, J. (1989). Prognostic significance of cytokeratin-positive breast cancer metastases. *Modern Pathology*, **2** , 516-520.

[57] Shirahata, S. (1975). Locally most powerful rank tests for independance with censored data. *The Annals of Statistics*, **3**, No. 1, 241-245.

[58] Skovgaard, I.M. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Prob.*, **24**, 875-887.

[59] Stablein, D., Carter, W. and Novak, J.(1981). The analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials*. **2**, 149-159.

[60] Tajuddin, I. H. (1994). Distribution-free test for symmetry based on Wilcoxon two-sample test. *Appl. Statist.*, **21**, 5, 409-416.

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

[62] Temme, N. M. (1982). The uniform asymptotic expansion of a class of integrals related to cumulative distribution functions. *SIAM J. Math. Anal.*, **12**, 239-253.

[63] 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.

[64] Van Der Waerden, B. L. (1953). Ein neuer test für das problem der zwei stichproen. *Math. Annalen*, **126**, 93-107.

[65] Van Eeden, C. and Benard, A. (1957). A general class of distribution free tests for symmetry containing the tests of Wilcoxon and Fisher. *Proceedings of the Koninklijke Nederlandsche Akademie van Wetenschappen*, **A60**, 381-408.

[66] Vorlickova, P. D. (1976). Rank tests for independance when samples are from noncontinuous distributions and censored. *Commentationes Mathematicae universitatis carolinae*. **17**, 3,557-565.

[67] Woolson R.F.(1981). Rank tests and a one-sample log-rank test for comparing obseved survival data to a standard population. *Biometrics*, **37**, 687-696.