Small Sample LD50 Confidence Intervals Using Saddlepoint Approximations

Robert L. Paige
Department of Mathematics and Statistics
Missouri University of Science and Technology
(formerly University of Missouri-Rolla)
Rolla, MO 65409
E-mail: paigero@mst.edu

Phillip L. Chapman
Department of Statistics
Colorado State University
Fort Collins, CO 80523
E-mail: pchapman@stat.colostate.edu

Ronald W. Butler
Department of Statistical Sciences
Southern Methodist Universiity
Dallas, TX 75275
E-mail: rbutler@smu.edu

September 29, 2010

Authors' Footnote:

Robert L. Paige is Associate Professor, Department of Mathematics and Statistics, Missouri University of Science and Technology (formerly University of Missouri - Rolla), Rolla, MO 65409 (E-mail: paigero@mst.edu). Phillip L. Chapman is Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523 (E-mail: pchapman@stat.colostate.edu). Ronald W. Butler is C.F. Frensley Professor of Mathematical Sciences, Department of Statistical Sciences, Southern Methodist University, Dallas, TX 75275 (E-mail: rbutler@smu.edu).

Abstract

Confidence intervals for the median lethal dose (LD50) and other dose percentiles in logistic regression models are developed using a generalization of the Fieller theorem for exponential families and saddlepoint approximations. Simulation results show that, in terms of one-tailed and two-tailed coverage, the proposed methodology generally outperforms competing confidence intervals obtained from the classical Fieller, likelihood ratio, and score methods. In terms of two-tailed coverage, the proposed method is comparable to the Bartlett-corrected likelihood ratio method, but generally outperforms it in terms of one-tailed coverage. An extension to the competing risk setting is presented that allows binary response adjustments to be made using observed censoring times.

KEYWORDS: Bartlett correction, binary data, bootstrap, competing risks, Fieller's method, likelihood ratio, saddlepoint approximation, score statistic.

1. INTRODUCTION

In toxicity experiments, subjects are given dose x, usually measured on a logarithmic scale, and a binary response (death or non-death) is observed. Let Y_i be the number of deaths among the n_i subjects receiving dose x_i, for $i = 1, \ldots I$. Define $LD100p$ as the dose for which the probability of death is $p \in (0, 1)$. When $p = 0.5$, the $LD100p$ is the median lethal dose, or $LD50$, which is commonly used as an overall measure of toxicity. The $LD100p$ for other values of p is also useful.

If subjects are assumed to be independent Bernoulli trials with probability of death π_i for subjects receiving dose x_i, then $Y_i \,|\, x_i \sim$ Binomial (n_i, π_i). The logistic model (McCullagh and Nelder, 1989) assumes that the logit of π_i is linear in dose so

$$\ln \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 x_i. \tag{1}$$

For fixed p, let $g := \ln\{p/(1 - p)\}$ and $\lambda := LD100p$ so the relationship is

$$g = \beta_0 + \beta_1 \lambda.$$

In a toxicity study it is reasonable to assume that $\beta_1 > 0$ so λ can be expressed as

$$\lambda = (g - \beta_0)/\beta_1.$$

The $LD100p$ is usually estimated by maximum likelihood, $\widehat{\lambda} = (g - \widehat{\beta}_0)/\widehat{\beta}_1$, where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the maximum likelihood estimates (MLEs) of the intercept and slope. An asymptotic standard error, $\widehat{\sigma}(\widehat{\lambda})$, can be based on a one-term Taylor series approximation of $(g - \widehat{\beta}_1)/\widehat{\beta}_2$ together with the asymptotic covariance matrix for $(\widehat{\beta}_0, \widehat{\beta}_1)^T$,

$$\widehat{V} = \begin{bmatrix} v_{00} & v_{01} \\ v_{01} & v_{11} \end{bmatrix},$$

which is computed by inverting the observed Fisher information.

There are a number of existing methods for constructing approximate confidence intervals (CIs) for $LD100p$. Some authors do not assume the logistic model at all but rather consider non-parametric methods for making inference about $LD100p$; see for instance Bhattacharya and Kong (2006), Dette et al. (2005), Glasby (1987) and Schmoyer (1984). Other authors such as Stukel (1988) consider parametric generalizations of the logistic model. We shall however consider methods where the logistic model is assumed since this is often what is done in practice.

The delta method is based on the approximate normality of $\widehat{\lambda}$. A $100(1-2\alpha)\%$ delta method CI is given by

$$\widehat{\lambda} \pm z_\alpha \widehat{\sigma}(\widehat{\lambda}),$$

where z_α is the upper α point of the standard normal distribution.

The Fieller CI (Fieller, 1954) is based on the asymptotic normality of the MLE, $(\widehat{\beta}_0, \widehat{\beta}_1)^T$. The function $(\widehat{\beta}_0 + \lambda\widehat{\beta}_1 - g)$ is approximately normally distributed with mean zero and estimated variance $v_{00} + 2\lambda v_{01} + \lambda^2 v_{11}$. The $100(1-2\alpha)\%$ Fieller interval is the set of λ satisfying the inequality

$$\frac{(\widehat{\beta}_0 + \lambda\widehat{\beta}_1 - g)^2}{v_{00} + 2\lambda v_{01} + \lambda^2 v_{11}} < z_\alpha^2. \tag{2}$$

Although Finney (1971) described the Fieller interval as a fiducial interval, it can be reasonably interpreted as an approximate CI. Indeed, if $(\widehat{\beta}_0, \widehat{\beta}_1)$ were exactly normal, and the variances were known, rather than estimated, the Fieller interval coverage would be exact.

A third CI for λ can also be formed as the set of λ not rejected by the asymptotic likelihood ratio (LR) test. This approach for the $LD50$ is advocated in Williams (1986), where it is noted that a LR interval can be computed even when the estimated slope or intercept is infinite, and the delta method and Fieller CIs cannot be computed. Williams also suggests that when computing LR CIs an upper bound be placed on the slope parameter, based on subjective area knowledge.

The parametric bootstrap method of Hwang (1995) is a refinement of Fieller's method. To construct a $100(1-2\alpha)\%$ CI for λ one needs to find, via simulation, q_α

and $q_{1-\alpha}$ the α and $1-\alpha$ quantiles of the distribution of $T_0^*(\widehat{\lambda})$ where

$$T_0^*\left(\lambda\right) = \frac{\widehat{\beta}_0^* + \lambda\widehat{\beta}_1^* - g}{\sqrt{v_{00}^* + 2\lambda v_{01}^* + \lambda^2 v_{11}^*}}$$

and the starred $(*)$ quantities denote parameter estimates obtained from data simulated from the fitted logistic model. The CI for λ consists of those values of λ such that

$$q_\alpha < T_0^*\left(\lambda\right) < q_{1-\alpha}.$$

The score CI consists of the set of λ satisfying the inequality

$$\frac{\sum_{i=1}^I \hat{\sigma}_\lambda^2\left(Y_i\right)\left(x_i-\lambda\right)^2\left[\sum_{i=1}^I\left\{y_i-\hat{\mu}_\lambda\left(Y_i\right)\right\}^2\right]}{\sum_{i=1}^I \hat{\sigma}_\lambda^2\left(Y_i\right)\left[\sum_{i=1}^I \hat{\sigma}_\lambda^2\left(Y_i\right)\left\{x_i-\tilde{x}_\lambda\right\}^2\right]} < z_\alpha^2. \tag{3}$$

where $\hat{\mu}_\lambda\left(Y_i\right)$ and $\hat{\sigma}_\lambda^2\left(Y_i\right)$ are the MLEs for the mean and variance, respectively, of Y_i under the constraint that λ is in fact the true value of $LD100p$, and $\tilde{x}_\lambda = \left\{\sum_{i=1}^I \hat{\sigma}_\lambda^2\left(Y_i\right)\right\}^{-1}\sum_{i=1}^I \hat{\sigma}_\lambda^2\left(Y_i\right)x_i$.

Finally, we also consider the CI obtained from the Bartlett-corrected likelihood ratio (BLR) test. The BLR test statistic is obtained by scaling the LR test statistic so it has a chi-squared distribution with error $O\left(n^{-2}\right)$; see Butler (2007, section 7.1.2) for more details. The scaling factor for $LD100p$ in the logistic model is given in Harris et al. (1999).

Although all of these interval methods apply to any $LD100p$, simulation studies have often focused on the $LD50$. Sitter and Wu (1993) compare the coverage rates of the first two methods and conclude that the Fieller method is superior by a wide margin. However, for small n_i, or for highly asymmetrical designs, the Fieller CIs have poor coverage. Faraggi, Izikson and Reiser (2003) compare the first three methods, as well as an adjusted likelihood method and a bootstrap ABC method. They conclude that for small samples the LR method has observed coverage that is closest to nominal, and that the Fieller method is generally conservative.

The BLR and score methods were introduced in Harris et al. (1999) where, for $LD50$, they were found to give a coverage probability closer to the nominal level than the Fieller and LR methods. Furthermore, Huang (2002) found that these methods also outperform the Fieller and LR methods for the estimation of LD90. Huang et al. (2000) considered a number of bootstrap and non-bootstrap methods for generating $LD90$ CIs and found that the score method generally performed very well in addition to the BC and BCa bootstrap methods. For $LD10$ and $LD50$ in the probit model, Mueller and Wang (1990) found that the percentile method and two variants, and the BC and BCa bootstrap CIs, exhibited no uniform improvement over the delta method. Despite the somewhat conflicting evidence in support of bootstrap methods for $LD100p$ problems, we consider the bootstrap method of Hwang (1995) since it has been found to work quite well in Fieller problems.

In this paper we propose a new method for $LD100p$ CIs based on a generalization of the Fieller theorem (Cox, 1967) for testing ratios of canonical parameters in exponential families. The new method involves transformation of the sufficient statistics so that the hypothesis $H_0 : \lambda = (g - \beta_0)/\beta_1 = \lambda_0$ is equivalent to the test that the first canonical parameter equals g. A uniformly most powerful similar test is then based on the distribution of the first sufficient statistic, given the value of the second sufficient statistic; see Lehmann (1986, sec. 4.4). The latter conditional distribution is approximated by the double-saddlepoint formula of Skovgaard (1987). One-sided $100(1-\alpha)\%$ CIs are formed as the set of λ_0 values not rejected in the one-sided hypothesis test. Two-sided $100(1-2\alpha)\%$ CIs are formed as the intersection of two one-sided intervals.

The new CI formula is derived in Section 2. An example using the Hewlett data set is given in Section 3. Simulation results for the $LD50$ and $LD90$ are given in Section 4. In a competing risks setting where censoring times are available, an extension of the methodology that adjusts for censoring and survival times is given in Section 5. Concluding remarks are included in Section 6.

2. SADDLEPOINT LD100p CI CONSTRUCTION

Ignoring constants that depend only on the data, the binomial likelihood is

$$\begin{aligned}\mathcal{L}(\pi_1, \ldots, \pi_I) &= \prod_{i=1}^{I} \pi_i^{y_i}(1-\pi_i)^{n_i - y_i} \\ &= \exp\left\{\sum_{i=1}^{I} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{I} n_i \ln(1-\pi_i)\right\}.\end{aligned}$$

After applying the logistic model,

$$\mathcal{L}(\beta_0, \beta_1) = \exp\left[\beta_0 \sum_{i=1}^{I} y_i + \beta_1 \sum_{i=1}^{I} y_i x_i - \sum_{i=1}^{I} n_i \ln\left\{1 + \exp(\beta_0 + \beta_1 x_i)\right\}\right].$$

In exponential family form the likelihood is written as

$$\mathcal{L}(\beta_0, \beta_1) = \exp\left\{\beta_0 d + \beta_1 r - c(\beta_0, \beta_1)\right\} \tag{4}$$

with sufficient statistics

$$d = \sum_{i=1}^{I} y_i \text{ and } r = \sum_{i=1}^{I} y_i x_i,$$

canonical parameter $\theta = (\beta_0, \beta_1)^T \in \Re^2$, and normalization constant

$$c\left(\beta_0, \beta_1\right) = \sum_{i=1}^{I} n_i \ln\left\{1 + \exp(\beta_0 + \beta_1 x_i)\right\}.$$

The MLE for the logistic model, $\widehat{\theta} = (\widehat{\beta}_0, \hat{\beta}_1)^T$, is obtained by solving the likelihood equations $c'\left(\beta_0, \beta_1\right) = (d, r)^T$, and asymptotic standard deviations are computed from $c''\left(\beta_0, \beta_1\right)^{-1}$, evaluated at the MLE. Formulas for $c'\left(\beta_0, \beta_1\right)$ and $c''\left(\beta_0, \beta_1\right)$ can be written as functions of π_i so

$$\pi_i = \pi_i(\beta_0, \beta_1) = \frac{\exp(\beta_0 + \beta_1 x_i\,)}{1 + \exp(\beta_0 + \beta_1 x_i\,)},$$

$$c'\left(\beta_0, \beta_1\right) = \begin{bmatrix} \sum_{i=1}^{I} n_i \pi_i \\ \sum_{i=1}^{I} n_i x_i \pi_i \end{bmatrix},$$

and

$$c''\left(\beta_0, \beta_1\right) = \begin{bmatrix} \sum_{i=1}^{I} n_i \pi_i\left(1-\pi_i\right) & \sum_{i=1}^{I} n_i x_i \pi_i\left(1-\pi_i\right) \\ \sum_{i=1}^{I} n_i x_i \pi_i\left(1-\pi_i\right) & \sum_{i=1}^{I} n_i x_i^2 \pi_i\left(1-\pi_i\right) \end{bmatrix}.$$

The large sample variance of the MLE is

$$\begin{bmatrix} v_{00} & v_{01} \\ v_{01} & v_{11} \end{bmatrix} := c''(\widehat{\beta}_0, \widehat{\beta}_1)^{-1}.$$

For a specified p and $g = \ln\{p/(1-p)\}$, and assuming $\beta_1 > 0$, consider the test of the hypothesis that $\lambda = \lambda_0$,

$$H_0 : \frac{g - \beta_0}{\beta_1} = \lambda_0,$$

versus the alternative

$$H_1 : \frac{g - \beta_0}{\beta_1} > \lambda_0.$$

In all that follows we omit the subscript on λ_0 to simplify the typesetting. For an hypothesized λ, define $\psi_\lambda = \beta_0 + \lambda\beta_1$, where the subscript λ is a reminder that the definition of ψ depends on the null hypothesis being tested. Hypotheses H_0 and H_1 above are equivalent to

$$H_{0\lambda} : \psi_\lambda = g \qquad \text{and} \qquad H_{1\lambda} : \psi_\lambda < g.$$

Adding and subtracting $\lambda\beta_1 d$ within the exponent of the likelihood we obtain

$$\begin{aligned} \mathcal{L}(\beta_0, \beta_1) &= \exp\{\beta_0 d + \lambda\beta_1 d + \beta_1 r - \lambda\beta_1 d - c(\beta_0, \beta_1)\} \\ &= \exp\{(\beta_0 + \lambda\beta_1) d + \beta_1(r - \lambda d) - c(\beta_0, \beta_1)\}. \end{aligned}$$

In the above expression, we recognize $(\beta_0 + \lambda\beta_1)$ as ψ_λ, define $z_\lambda := (r - \lambda d)$, and substitute $\psi_\lambda - \lambda\beta_1$ for β_0 in the function $c(\cdot, \cdot)$ to obtain

$$\mathcal{L}(\beta_0, \beta_1) = \exp\{\psi_\lambda d + \beta_1 z_\lambda - c(\psi_\lambda - \lambda\beta_1, \beta_1)\}.$$

The likelihood is now in canonical form for the parameter (ψ_λ, β_1) and sufficient statistics (d, z_λ). The transformation from (d, r) to (d, z_λ) is one-to-one and can be written explicitly as

$$\begin{pmatrix} d \\ z_\lambda \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ -\lambda & 1 \end{bmatrix} \begin{pmatrix} d \\ r \end{pmatrix} := B_\lambda \begin{pmatrix} d \\ r \end{pmatrix}.$$

The transformation of the parameters is the transpose of the inverse transformation

$$\begin{pmatrix} \psi_\lambda \\ \beta_1 \end{pmatrix} = \begin{bmatrix} 1 & \lambda \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \left(B_\lambda^{-1}\right)^T \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

The transformation matrices have determinant equal to 1, which is not a consideration for a likelihood written with respect a discrete measure, but is relevant to the continuous approximation used below.

In the new parameterization, the hypothesis $H_{0\lambda}$ is a point hypothesis about the first canonical parameter. The UMP unbiased level-α test for testing $H_{0\lambda} : \psi_\lambda = g$ versus $H_{1\lambda} : \psi_\lambda < g$ is the conditional test based on the distribution of d given z_λ. (Lehmann, 1986, sec. 4.4). In the continuous case, the similar conditional test rejects $H_{0\lambda}$ when

$$P(D < d \,|\, Z_\lambda = z_\lambda : \psi_\lambda = g) < \alpha.$$

Because the distribution is an exponential family, this conditional distribution does not depend on the nuisance parameter β_1. For discrete distributions, a randomization mechanism must be employed to achieve exact conditional level α. This problem is described by Cox (1967) in the context of his generalization of Fieller's Theorem for exponential families. Cox described the properties of the analogous conditional test for the ratio of Poisson means, and noted that for small sample sizes the test was not useful because the reference set for the conditional inference is very limited, or even degenerate. By analogy, Cox dismissed the conditional approach for such tests concerning the $LD50$ in logistic regression. In the $LD50$ problem, the reference set is the set of outcomes for which $r - \lambda d$ equals its observed value. This set is indeed very limited, and even degenerate for many real values of λ. The conditional p-value may be small for one value of λ, and yet large for nearby values of λ. If confidence regions are formed as the set of values of λ not rejected in the corresponding hypothesis test, such regions will not even be intervals.

Rather than dismiss the conditional test, we proceed by computing the Skovgaard (1987) double saddlepoint approximation to the conditional distribution: $D \,|\, Z_\lambda =$

$z_\lambda : \psi_\lambda = g$. Though the Skovgaard approximation is derived for continuous distributions, we apply it formally to the discrete case, *without correction for continuity.* One justification for this approach, given by Davison and Wang (2002), is that the saddlepoint approximation is nearly an exact solution for an analogous problem in which the discrete mass function of the data is replaced by a continuous approximation. Another justification is given by Pierce and Peters (1999), who argue in favor of approximate conditioning in discrete problems. They note that the p-value resulting from saddlepoint approximation accomplishes approximate conditioning, and provides an approximation to a conditional mid-p-value. The use of mid-p-values when forming CIs based on discrete distributions is advocated in Agresti (1992) and Routledge (1994). Ultimately, however, justification for our approach rests upon the coverage accuracy of the interval and its improvement over existing methods. Saddlepoint methods have achieved remarkable accuracy in approximating non-normal distributions (Butler, 2007) and potentially offer improvement over existing methods in very small samples or asymmetric situations where methods based on normal approximations fail.

Define $F(d \,|\, z_\lambda) := \Pr(D \leq d \,|\, Z_\lambda = z_\lambda : \psi_\lambda = g)$. The Skovgaard approximation to $F(d \,|\, z_\lambda)$ for exponential families, also described in Butler (2007, sec. 5.4.5), is

$$\widehat{F}(d \,|\, z_\lambda) := \Phi(\widehat{w}_\lambda) + \phi(\widehat{w}_\lambda)\left(\frac{1}{\widehat{w}_\lambda} - \frac{1}{\widehat{u}_\lambda}\right), \qquad \widehat{\psi}_\lambda \neq g \tag{5}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal CDF and PDF, respectively, and $\widehat{w}_\lambda$ and $\widehat{u}_\lambda$ are

$$\widehat{w}_\lambda := \operatorname{sgn}(\widehat{\psi}_\lambda - g)\sqrt{2\ln\mathcal{L}(\widehat{\beta}_0, \widehat{\beta}_1) - 2\ln\mathcal{L}(g - \lambda\widehat{\beta}_{1\lambda}, \widehat{\beta}_{1\lambda})}, \tag{6}$$

$$\widehat{u}_\lambda := (\widehat{\psi}_\lambda - g)\sqrt{\frac{|c''(\widehat{\beta}_0, \widehat{\beta}_1)|}{\widehat{h}_\lambda}},$$

where

$$\widehat{h}_\lambda = \begin{pmatrix} -\lambda & 1 \end{pmatrix} c''(g - \lambda\widehat{\beta}_{1\lambda}, \widehat{\beta}_{1\lambda})\begin{pmatrix} -\lambda \\ 1 \end{pmatrix}.$$

The $\widehat{h}_\lambda$ term is the second derivative of $c(\psi_\lambda - \lambda\beta_1, \beta_1)$ w.r.t. β_1, evaluated at $(g - \lambda\widehat{\beta}_{1\lambda}, \widehat{\beta}_{1\lambda})$, the MLE under the constraint $H_{0\lambda}$. $\widehat{\beta}_{1\lambda}$ solves likelihood equation

$$\sum_{i=1}^{I} (x_i - \lambda) \left[\frac{n_i \exp\{g + \widehat{\beta}_{1\lambda}\,(x_i - \lambda)\}}{1 + \exp\{g + \widehat{\beta}_{1\lambda}\,(x_i - \lambda)\}} - y_i \right] = 0.$$

The formula for $\widehat{w}_\lambda$ is a signed version of the square root of the usual likelihood ratio statistic for testing $H_{0\lambda}$ versus $H_{1\lambda}$, and the leading term in the expression for $\widehat{F}(d \mid z_\lambda)$ is the one-sided p-value for the usual likelihood ratio test.

One-sided CIs are formed using the values of λ not rejected by the size α test. Thus, a $100(1-\alpha)\%$ lower one-sided confidence limit is

$$LCL = \inf\{\lambda : \widehat{F}(s \,|\, z_\lambda) > \alpha\}.$$

Similarly, a $100(1-\alpha)\%$ upper one-sided confidence limit is formed using the values of λ not rejected in the test of $H_{0\lambda}$ above, versus $H_{1\lambda} : \psi_\lambda > g$:

$$UCL = \sup\{\lambda : 1 - \widehat{F}(s \mid z_\lambda) > \alpha\}.$$

The two-sided interval is the intersection of the two one-sided intervals. An unusual characteristic of this CI construction is the dependence of the conditioning variable z_λ on λ.

For fixed λ, individual p-values are easily computed from quantities that are standard output of most logistic regression computer packages. The quantity $\widehat{w}_\lambda$ is the signed square root of the likelihood ratio statistic for testing $H_{0\lambda}$ above, versus $H_{1\lambda} : \psi_\lambda > g$. It involves the difference between the deviances of the unrestricted and null models. The null model may be fitted in a statistical package by transforming x_i to $x_i - \lambda$, and specifying a model without intercept and with offset value equal to g. The quantity $\widehat{u}_\lambda$ is the MLE of $\widehat{\psi}_\lambda$, minus its hypothesized value, divided by its asymptotic standard error. The value $c''(\widehat{\beta}_0, \widehat{\beta}_1)$ is the inverse of the estimated covariance matrix of the unrestricted parameter estimates, under the alternative model. The value $\widehat{h}_\lambda$ is the inverse of the estimated variance of the $\widehat{\beta}_{1\lambda}$ parameter in the null model.

To compute the $100(1-2\alpha)\%$ CI, the p-value must be computed over a fine grid, or inserted into a root-finding subroutine to find the values of λ at which $\widehat{F}(d\,|\,z_\lambda)$ equals α and $1-\alpha$. The likelihood ratio CI is obtained by the same procedure, using only the leading term in $\widehat{F}(d\,|\,z_\lambda)$.

3. EXAMPLE : HEWLETT DATA

In this section we examine 95% CIs for the Hewlett data using the saddlepoint, Fieller, LR, Hwang's bootstrap, score and BLR methods. We do not consider the delta method due to its poor performance in small to moderate sample sizes. The Hewlett data, with nine design points, relatively moderate samples, and a steep response curve is studied in Sitter and Wu (1997), and Faraggi *et al.* (2003). It seems to have first appeared in print in Abdelbasit and Plackett (1983) where the authors state that it was obtained from personal communication with P.S. Hewlett; no additional details were provided. Based upon Hewlett (1947,1975) we posit that some kind of petroleum oil was sprayed directly on each of the grain weevils in the study. It appears that this oil caused death by migrating into the tracheae and causing anoxia (a total decrease in the level of oxygen). Furthermore, death, when it occurred, probably happened within a few days of the spraying and those weevils which did not die recovered from the spraying. The data are given in Table 1.

The quantities needed for computing $\widehat{F}(d\,|\,z_\lambda)$ are easily obtained from logistic regression programs. The unconstrained MLEs, $\widehat{\beta}_0 = 0.4891892$, $\widehat{\beta}_1 = 28.2422$, and the deviance, $-2\ln\mathcal{L}(\widehat{\beta}_0,\widehat{\beta}_1) = 22.78897$, are standard output. For $LD50$ intervals, set $p = 0.5$ and $g = 0$. Fix $\lambda = -0.01$ and compute $\widehat{\psi}_\lambda = \widehat{\beta}_0 + \lambda\widehat{\beta}_1 = 0.2067674$. To compute the MLE and deviance under the constraint, $g = \beta_0 + \lambda\beta_1$, we note that the constraint implies that the linear predictor is $(g-\lambda\beta_1)+\beta_1 x_i = g+\beta_1(x_i-\lambda)$, a no-intercept model with independent variable $x_i - \lambda$ and offset g. From any logistic regression program one can obtain $\widehat{\beta}_{1\lambda} = 27.2273$ and deviance $-\ln\mathcal{L}(g-\widehat{\beta}_{1\lambda}\lambda,\widehat{\beta}_{1\lambda}) =$

23.59636. Thus $\widehat{w}_\lambda = 0.898548$. Computing $\widehat{u}_\lambda = 0.848501$ requires evaluating $c''(.,.)$ under the unconstrained and unconstrained parameter estimates, but only depends on the parameter estimates through the model predicted values, $\widehat{\pi}_i$. Thus $\widehat{F}(d \,|\, z_\lambda) = 0.798064$. Figure 1 plots $\widehat{F}(d \,|\, z_\lambda)$ versus λ for λ in $[-0.06, 0.02]$ and identifies the two-sided 95% CI as the values of λ for which $0.025 < \widehat{F}(d \,|\, z_\lambda) < 0.975$.

Estimates of $LD10$, $LD50$, and $LD90$, and their respective 95% CIs, calculated using the six methods, are given in Table 2. The differences between the CI methods are very minor, due to the relatively moderate sample size; however, the saddlepoint CI is generally narrower than the Fieller and the score CIs which are similar, and is about the same width as the LR and BLR CIs, and Hwang's CI based on 1000 bootstrap samples.

4. SIMULATION STUDIES AT HEWLETT DESIGN POINTS

In this section we compare the performance of the saddlepoint CIs, for $LD50$ and $LD90$, with CIs from the Fieller, LR, score and BLR methods. The resampling-based methodology of Hwang (1995) was excluded from these simulation studies since it was found to have substantial undercoverage for small samples and highly-skewed designs in a preliminary simulation study. In this preliminary study and the main study, the Hewlett design points are chosen so as to be comparable to the simulations of Sitter and Wu (1993) and Faraggi *et al.* (2003). For the LD50 simulations, we take $\beta_1 = 7, 14$ and 21 and $\lambda = -0.017$ (MLE for LD50), 0.1, 0.2 and 0.3, and sample size $n = 7, 10, 20, 30$ and 50. Here sample size is the number of independent realizations simulated at each Hewlett design point. The λ grid does not contain negative λ values -0.1,-0.2 or -0.3 since symmetry guaranteed that results for these values would match the results of their positive λ counterparts up to simulation error, where the roles of death and non-death are interchanged.

The λ grid for the LD90 simulations consists of values 0.0, 0.1, 0.2, and 0.3. The MLE for LD90 is 0.0605 and is too close to 0.1 to be of interest so $\lambda = 0.0$ is used instead. In addition, negative λ grid values are not considered since they will yield many simulated data sets with very few unaffected subjects and hence will provide very little information about LD90. We do not consider simulations for LD10 since this dose level for deaths is the LD90 dose level for non-deaths and the Hewlett design points are symmetric about zero. As such, results for LD10 would coincide with those for LD90 if the roles of death and non-death are interchanged. The saddlepoint, Fieller, score and BLR methods cannot be computed in the presence of infinite MLEs and as such we let these methods default to the LR method in such settings.

For the preliminary simulation study we followed the simulation design of Hwang (1995) and performed 3,000 Monte Carlo simulations, each involving 1000 bootstrap samples. Here the coverage probability was averaged over 21 equi-spaced λ values, in an interval containing the true value of λ, for reasons we explain below. Table 3 displays results for those cases where the coverage probability for a nominal 95% CI was particularly bad (below 0.9) for smaller samples. Notice that the observed coverage probabilities were all relatively close to the nominal value of 0.95 for larger n whereas they were quite poor for small n. In the main simulation study none of the five remaining methods ever yielded 95% CIs with coverage probabilities below 0.9.

For the five remaining methods we define several types of CI error. A low error occurs when the CI lies completely below λ and a high error occurs if it lies completely above this value. Low and high error rates exhibit oscillatory behavior when graphed over a range λ values. Figure 2 shows low error rates for the saddlepoint (black), Fieller (red) LR (green), score (blue) and BLR (gold) methods, for $\beta_1 = 7$ and $n = 10$, where 100,000 data sets were simulated at each point of a fine grid of LD50 λ values. The oscillatory behavior seen in this figure is caused by the discrete nature of the underlying binomial distribution. A similar phenomenon is noted for instance in Brown, Cai and DasGupta (2002) for the binomial proportion exact CIs considered therein. We adjust for error rate oscillation by considering averaged error rates where the average is taken over 21 equi-spaced λ values in the interval $[\lambda - 0.01, \lambda + 0.01]$.

Furthermore, the error rate at each of the λ grid points will be based on 50,000 Monte Carlo simulations for a total of 1,050,000 data sets per configuration and simulation standard error of 0.015% when the true error rate is 2.5%.

We also consider median CI length based upon 50,000 Monte Carlo repetitions at the specified value of λ. Numerical experimentation revealed no oscillation for this summary statistic so averaging over the grid of λ values was not necessary.

For smaller samples, all five of the remaining methods sometimes produced CIs that were infinite in length, however this is not necessarily a sign that the methods are failing. In many cases it is an indication that the data do not contain enough information about the slope of the logistic curve to make useful conclusions about the location of λ, e.g. indicating that more data should be collected. Hwang (1995) points out that in Fieller problems a CI method that always has a positive confidence level must on occasion produce infinite CIs. Furthermore, with regards to $LD100p$, an anonymous referee indicated to us that with β_0 fixed and β_1 approaching zero the coverage probabilities for any method, which always produces finite intervals, will have to approach zero; see Theorem 1 of Gleser and Hwang (1987).

4.1 $LD50$ Results

We first compare the performance of the saddlepoint CI for $LD50$ with CIs from the Fieller, LR, score and BLR methods. In the figures below "Combined Error" refers to the two-tailed error rate which is the sum of the low and high error rates. Furthermore, we report averaged error rates. From Figures 3 and 4 one can see that the saddlepoint (SP) method outperforms the Fieller and LR methods by a wide margin and the score and BLR methods to a lesser extent. The saddlepoint method is nearly perfect in terms of one-tailed and two-tailed coverage, for all cases considered. The Fieller method is conservative in terms of two-tailed coverage and the LR method is liberal, as is the score method. The BLR has good two-tailed coverage but much poorer one-tailed coverage. While there appears to be relatively few differences in the plots of median lengths, the score method seems to generally have the shortest median length of all methods which is consistent with its liberal coverage.

Another issue to consider in the simulations is that Fieller's method yields a finite CI whenever the slope is statistically significant, i.e. $\hat{\beta}_1^2/v_{11} > z_\alpha^2$. It is argued in Sitter and Wu (1993) that the LD50 estimation is not meaningful when the regression relationship is not significant and therefore one should consider CI error rates which are defined over the collection of data sets which result in a finite Fieller CI. We refer to these error rates as Fieller-conditional. Previous studies have reported only Fieller-conditional error rates, but in doing so have sometimes omitted over half of the simulated data sets.

For the settings considered in Figure 3, the Fieller-conditional error rates and unconditional error rates are essentially the same since the proportion of cases with infinite Fieller CIs was so small. As a result, we have presented only one set of figures. Figure 5 presents Fieller-conditional error rates for the same settings as those used in Figure 4 as well as the percentage of times Fieller's method did not yield a finite CI. Inspection of this figure reveals that with few exceptions, in terms of Fieller-conditional one-sided coverage, the saddlepoint method outperforms the BLR method. In terms of both one-sided and two-sided coverage, it outperforms the other methods with few exceptions. These exceptions occur with large λ and or β_1 which are cases in which the design is poorly suited to provide much information about λ and therefore ones in which any method of CI construction will perform poorly. For instance, with settings $\lambda = 0.3$, $\beta_1 = 21$ and $n = 7$ the median confidence interval length for all methods was infinite and so are not given in Figure 4. The reason for this is seen in Figure 5 where the percentage of infinite Fieller intervals is shown to be 65%.

A referee has pointed out the need to assess the bias associated with defaulting to the LR confidence interval in the case of infinite MLE's. To assess this bias we consider MLE-conditional CI error rates which are defined as rates over the collection of data sets which result in finite MLEs. We take as our measure of bias the difference of two Euclidean distances. The first distance is that of the unconditional error rate from the nominal rate and the second is that of the MLE-conditional error rate from the same nominal rate. Our bias measure describes how much LR defaulting improves the error rates for a CI method. A positive bias value indicates that LR defaulting improved the error rate and a negative bias value indicates that it actually hurt the error rate. For the larger samples, $n = 20, 30$ and 50, the bias values are essentially zero so these graphs are omitted. Figure 6 indicates that LR defaulting has the greatest effect for larger values of λ and or β_1 and that the saddlepoint error rates are relatively unaffected by LR defaulting. Furthermore, the bias due to LR defaulting is greatest for all methods when $\beta_1 = 21$ which coincides with higher rates of infinite MLEs.

4.2 $LD90$ Results

In this section we compare the performance of the saddlepoint CI's for $LD90$ with CIs from the Fieller, LR, score and BLR methods. For the larger sample sizes, the proportion of cases where the Fieller CI was infinite or the MLEs were infinite did not affect the results so we present only the unconditional error rates. As before, there was a difference in the unconditional, Fieller-conditional and MLE-conditional error rates for smaller samples so we report Fieller-conditional error rates and LR default bias rates for these settings.

From Figures 7 and 8, it can be seen that the saddlepoint method is again nearly perfect in terms of one-tailed and two-tailed coverage for all cases considered with few exceptions. The exceptions again occur with small n and large λ and or β_1, cases for which any method would perform poorly. As before, the saddlepoint method generally outperforms the Fieller and LR methods by a wide margin and the score and BLR methods to a lesser extent. The Fieller method is again conservative in terms of two-tailed coverage and the LR method is liberal, as is the score method. A comparison of Figure 3 with Figure 7 and Figure 4 with Figure 8 shows that Fieller's method is more conservative for LD90 than for LD50 and similarly the LR and score methods are more liberal for LD90 than for LD50. The BLR method generally has good two-tailed coverage but it's LD90 one-tailed coverage is generally worse than its LD50 one-tailed coverage. As before, there appears to be relatively few differences in the plots of median lengths and the score method seems to have the shortest median length.

Figure 9 presents Fieller-conditional error rates for the same settings as those considered in Figure 8 as well as the percentage of times Fieller's method did not yield a finite CI. Inspection of this figure reveals that again with few exceptions, in terms of Fieller-conditional one-sided coverage, the saddlepoint method continues to outperform the BLR method. In terms of both one-sided and two-sided coverage, it outperforms the other methods with few exceptions. These exceptions as usual occur with large λ and or β_1.

The bias values for LR defaulting are shown in Figure 10 which indicates that LR defaulting again has the greatest effect for larger values of λ and or β_1 and that the saddlepoint error rates are relatively unaffected by LR defaulting in comparison with the other methods.

5. SURVIVAL-ADJUSTED $LD100p$ CONFIDENCE INTERVAL

It has been pointed out by an Associate Editor that a dose-response analysis in a toxicity or carcinogenicity study often needs to address a competing risk problem. For instance with the Hewlett data, if the time between the application of oil and death had not been so short (2-3 days), then a grain weevil could also have died from natural causes before dying from the dosing or else could have been right-censored. For such applications, appropriate competing risk models for survival times are introduced along with the possibility of independent right censoring. When no censoring occurs, survival times are not found to be informative about $LD100p$ and the analysis based strictly on dose responses is fully informative about $LD100p$. If survival times are censored however, then such censoring times along with all known survival times carry additional information about $LD100p$. A method that incorporates some of this information into the present logistic regression analysis is described. It essentially involves adjusting death counts Y_i and sample sizes n_i at dose x_i by imputing the missing binary dose responses due to censoring. This scheme would increase Y_i and n_i to Y_i^* and n_i^* for $i = 1, 2, \ldots, I$.

To motivate such imputations, consider a semi-Markov competing risk model. The overall survival distribution is the sum of two subdistributions or cumulant incident functions (CIFs)

$$F(t|x;\beta) = p(x;\beta)F_1(t|x;\beta) + \{1 - p(x;\beta)\}F_2(t|x;\beta), \qquad (7)$$

where the first term is toxin (or radiation) specific and the second is specific to all other risks. Here, $p(x;\beta)$ is the probability that an animal ultimately dies from a toxin (or radiation) and $F_1(t|x;\beta)$ is the survival distribution of the animal if it is certain to die from the toxin and not from a competing cause. Survival distribution $F_2(t|x;\beta)$ applies to all other competing risks. In most examples, one would expect that $F_1(t|x;\beta) \neq F_2(t|x;\beta)$ are radically different however these distributions are the same in an important special case.

5.1 Markov Setting

Suppose $F_1(t|x;\beta)$ is Exponential (λ_1) with

$$\ln \lambda_1 = \lambda_{11} + \lambda_{12}x$$

and independent of $F_2(t|x;\beta)$ which is Exponential (λ_2). Then

$$\begin{aligned} p(x;\beta) &= \int_0^\infty F_1(t|x;\beta)dF_2(t|x;\beta) = \int_0^\infty (1-e^{-\lambda_1 y})\lambda_2 e^{-\lambda_2 y}dy \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{\exp(\lambda_{11}-\ln\lambda_2+\lambda_{12}x)}{\exp(\lambda_{11}-\ln\lambda_2+\lambda_{12}x)+1} \\ &= \frac{\exp(\beta_0+\beta_1 x)}{\exp(\beta_0+\beta_1 x)+1} = \frac{e^g}{e^g+1}, \end{aligned}$$

where $\beta_0 = \lambda_{11} - \ln\lambda_2$ and $\beta_1 = \lambda_{12}$. This model leads to logit probabilities for death by the toxin.

Since the competing risk is among independent exponentials, the CIF components $F_1(t|x;\beta) = F_2(t|x;\beta)$ are Exponential $(\lambda_0+\lambda_1)$ and

$$F(t|x;\beta) = F_1(t|x;\beta) = F_2(t|x;\beta). \tag{8}$$

The main consequence of (8) is that observed survival times, and right censoring times for which cause of death is unobserved, are uninformative about associated dose response. Thus in this Markov setting, observed death times by either cause and independently right-censoring times add no additional information for determining $LD100p$ and should therefore be ignored. To see this, denote dose, death time, and cause of death as $\{(x_i,t_i,z_i) : i = 1,\ldots,n\}$ where $z_i = 1$ $(z_i = 0)$ indicates death by toxic dose (competing risk). Denote independently right censored data as $(\mathfrak{x}_j,\mathfrak{t}_j) : j = 1,...,m$ where $\mathfrak{x}_j$ is dose, $\mathfrak{t}_j$ is censoring time, and binary toxic response $\mathfrak{z}_j$ is unobserved. In the Markov case, the likelihood function is

$$\begin{aligned} &\prod_{i=1}^{n} \{p(x_i;\beta)dF_1(t_i|x_i;\beta)\}^{z_i}\left\{[1-p(x_i;\beta)]\,dF_1(t_i|x_i;\beta)\right\}^{1-z_i} \times \prod_{j=1}^{m} S_1(\mathfrak{t}_j|\mathfrak{x}_j;\beta)\} \\ &= L_M(\beta) \times h(\mathfrak{t};F_1), \end{aligned}$$

where $S_1(\mathfrak{t}_j|\mathfrak{x}_j;\beta) = 1 - F_1(\mathfrak{t}_j|\mathfrak{x}_j;\beta)$, and

$$\begin{aligned} L_M(\beta) &= \prod_{i=1}^{n} \{p(x_i;\beta)\}^{z_i} \left[1 - p(x_i;\beta)\right]^{1-z_i} \\ h(\mathbf{t};F_1) &= \prod_{i=1}^{n} dF_1(t_i|x_i;\beta) \times \prod_{j=1}^{m} \{1 - F_1(\mathfrak{t}_j|\mathfrak{x}_j;\beta)\} \\ &= \tau^n \exp\left\{-\tau \left(\textstyle\sum_{i=1}^{n} t_i + \sum_{j=1}^{m} \mathfrak{t}_j\right)\right\} \prod_{i=1}^{n} dt_i, \end{aligned}$$

with $\tau = \lambda_0 + \lambda_1$. The likelihood is completely separable into logistic regression factor $L_M(\beta)$, with parameters $\beta = (\beta_0, \beta_1) = (\lambda_{01} + \ln \lambda_1, \lambda_{01})$, and distribution parameter $F_1 \leftrightarrow \tau = \lambda_0 + \lambda_1$ with factor $h(\mathbf{t};F_1)$. The likelihood is variation independent in β and τ; see Butler (2007, §9.6). Term $L_M(\beta)$ is the marginal likelihood function as it involves only uncensored dose response data and represents the marginal mass function of $\{z_i\}$. It does not represent the conditional distributions of $\{t_i|z_i\}$ for $i = 1, ..., n$ nor the distributions of the $\{\mathfrak{t}_j\}$. The separation of parameters in the likelihood is the strongest possible and likelihood inference under such circumstances should be based only on the marginal likelihood $L_M(\beta)$ as in the first part of this paper.

5.2 Non-Markov Setting

A special case of the semi-Markov mixture model in (7) is equivalent to the mixture model in Larson and Dinse (1985). It represents a more realistic non-Markov setting in which $F_1(t|x;\gamma) \neq F_2(t|x;\gamma)$ and γ is a nuisance parameter that does not include β. Without censoring, the likelihood is again variation independent in β and γ and logistic regression in observed dose responses is fully informative about $LD100p$. However, when right-censoring occurs, censoring times contribute likelihood factors that link together interest parameters β and nuisance parameters γ, F_1, and F_2.

Denoting marginal survival as $S(\mathfrak{t}_j|\mathfrak{x}_j;\beta,\gamma) = 1 - F(\mathfrak{t}_j|\mathfrak{x}_j;\beta,\gamma)$ and $F_k'(t_i|x_i;\gamma) = f_k(t_i|x_i;\gamma)$, the likelihood is

$$\prod_{i=1}^{n} \{p(x_i;\beta) f_1(t_i|x_i;\gamma)\}^{z_i} \{[1-p(x_i;\beta)]\, f_2(t_i|x_i;\gamma)\}^{1-z_i} \times \prod_{j=1}^{m} S(\mathfrak{t}_j|\mathfrak{x}_j;\beta,\gamma)$$
$$= L_M(\beta) \times \prod_{i=1}^{n} f_1(t_i|x_i;\gamma)^{z_i} f_2(t_i|x_i;\gamma)^{1-z_i} \times \prod_{j=1}^{m} S(\mathfrak{t}_j|\mathfrak{x}_j;\beta,\gamma).$$

Without the last factor due to censored data, the likelihood is completely separable into interest parameter β and nuisance parameters γ, f_1, and f_2. Marginal likelihood $L_M(\beta)$ is fully informative about β and the conditional densities for $\{t_i|\, z_i\}$ are fully informative about γ. When there is censored data, the last factor $\prod_{j=1}^{m} S(\mathfrak{t}_j|\mathfrak{x}_j;\beta,\gamma)$ involves both β and $(\gamma,\, f_1, f_2)$ and is the marginal distribution of $\{\mathfrak{t}_j\}$ that sums out the missing binary toxic responses $\{\mathfrak{z}_j : j = 1, \ldots, m\}$.

With censored data, inference would benefit from some alternative method that uses the information in $\{\mathfrak{t}_j; j = 1, \ldots, m\}$. A reasonable and practical solution is to impute the missing binary toxic responses $\{\mathfrak{z}_j : j = 1, \ldots, m\}$ in the censored data by using the EM algorithm (Dempster et al., 1977). Imputed data pairs $\hat{\mathcal{D}} =\{(\mathfrak{x}_j, \hat{\mathfrak{z}}_j) : j = 1, \ldots, m\}$ can then supplement data $\mathcal{D}=\{(x_i, z_i) : i = 1, \ldots, n\}$ expressed in marginal likelihood $L_M(\beta)$. New sample size $n + m$ accommodates both zero-one dose responses $\{z_i\}$ as well as imputed fractional dose responses $\{\hat{\mathfrak{z}}_j \in (0, 1)\}$. A "pseudo-likelihood" of the form (4) with canonical parameter $\theta = (\beta_0, \beta_1)^T \in \Re^2$ uses pseudo-sufficient statistics

$$\hat{d} = \sum_{i=1}^{n} z_i + \sum_{j=1}^{m} \hat{\mathfrak{z}}_j \text{ and } \hat{r} = \sum_{i=1}^{n} x_i z_i + \sum_{j=1}^{m} \mathfrak{x}_j \hat{\mathfrak{z}}_j,$$

and normalization constant

$$\hat{c}\,(\beta_0, \beta_1) = \sum_{i=1}^{n} z_i \ln\{1 + \exp(\beta_0 + \beta_1 x_i)\} + \sum_{j=1}^{m} \hat{\mathfrak{z}}_j \ln\{1 + \exp(\beta_0 + \beta_1 \mathfrak{x}_j)\}\,.$$

Such pseudo-likelihoods are of relevance only because they represent a model that motivates the inputs used in the saddlepoint formulas based on "data" $\mathcal{D} \cup \hat{\mathcal{D}}$.

Imputation $\hat{\mathfrak{z}}_j$ makes use of Bayes theorem and the fact that $S_1(t|x;\gamma)$ and $S_2(t|x;\gamma)$ are different distributions. For an observed value $\mathfrak{t}_j$, the imputation step of EM uses the Bayes update

$$\begin{aligned}\hat{\mathfrak{z}}_j^{(k+1)} &= P\left(\text{death by toxin}|t > \mathfrak{t}_j; \hat{\beta}^{(k)}, \hat{\theta}^{(k)}\right) = P\left(\mathfrak{z}_j = 1|t > \mathfrak{t}_j; \hat{\beta}^{(k)}, \hat{\theta}^{(k)}\right) \\ &= \frac{p(\mathfrak{x}_i; \hat{\beta}^{(k)})\hat{S}_1^{(k)}(\mathfrak{t}_j|\mathfrak{x}_j; \hat{\gamma}^{(k)})}{\hat{S}^{(k)}(\mathfrak{t}_j|\mathfrak{x}_j; \hat{\gamma}^{(k)})}\end{aligned}$$

where $\hat{\beta}^{(k)}$ and $\hat{\theta}^{(k)} = (\hat{\gamma}^{(k)},\hat{S}_1^{(k)}, \hat{S}_2^{(k)})$ are MLEs based on "complete" data $\{(x_i, t_i, z_i) : i = 1, \ldots, n\} \cup \{(\mathfrak{x}_j, \mathfrak{t}_j, \hat{\mathfrak{z}}_j^{(k)}) : j = 1, \ldots, m\}$. Then $\hat{\mathfrak{z}}_j$ is the limit point for $\{\hat{\mathfrak{z}}_j^{(k)}\}$.

Note that using imputed values of $\{\hat{\mathfrak{z}}_j\}$ is not the exact maximum likelihood procedure. That procedure would compute the $LD100p$ interval by entertaining the test $H_0 : (g - \beta_0)/\beta_1 = \lambda$ versus $H_1 : (g - \beta_0)/\beta_1 \neq \lambda$ and incorporating the missing data structure into each individual test of λ. Since the missing data structure is not of exponential form, this cannot be easily done. However our missing data approach is an ad hoc way of approximating this procedure.

5.3. Example

We illustrate the proposed non-Markov methodology with the data from Groer (1978) which describes the incidence of osteosarcomas in beagle dogs injected with varying amounts of Plutonium-239 (Ci/kg). Dogs in the study were injected with this carcinogen at dates between January 12, 1952 and October 17, 1974. Here 199 dogs were observed until death, to determine the number of days since injection, or until March 31, 1977 which was the cutoff date for data as originally reported in Jee (1977). At the time of death, an autopsy was performed to determine cause of death. Groer (1978) does not report survival/censoring times at lower dose values where many of the dogs were still alive at the cutoff date. We obtained these times from Jee (1977) and in Table 4 provide a summary, by mean dose level, of the number of dogs, the number of deaths, the number dead dogs with osteosarcoma and the number of censored death times. We omit the death and censoring times, which range from 467 to 5362. Our data set differs in a single respect from that used in Groer (1978); we maintain a cutoff date of March 31, 1977 and so do not take into account the two later osteosarcomas referenced in table footnote “*” on page 4089 of Groer (1978). Furthermore, in our analysis we exclude the data for all dogs at mean dose value 2.88 Ci/kg since the logistic model fits the remaining data well with a Hosmer-Lemeshow p-value of 0.321.

We assume a proportional hazards model using piecewise-exponential baseline cause-specific hazard functions, with $q+1$ subintervals,

$$\lambda_j\left(t|x,\gamma\right) = \exp\left(\gamma_{q+1,j}x\right)\sum_{i=0}^{q}\gamma_{ji}1\left[\tau_i \leq t < \tau_{i+1}\right) \tag{9}$$

where $j = 1$ or 2, $0 = \tau_0 < \tau_1 < \cdots < \tau_q < \tau_{q+1} = \infty$ and

$$1\left[\tau_i \leq t < \tau_{i+1}\right) = \begin{cases} 1 & \text{if} \quad \tau_i \leq t < \tau_{i+1} \\ 0 & \qquad \text{elsewhere} \end{cases} \quad .$$

The piecewise-exponential hazard model provides a simple and flexible distribution for modeling individual subdistributions or CIFs; see for example Proschan and Kim (1989). Larson and Dinse (1985) also assume the piecewise-exponential proportional hazards for the methodology they develop. Note that in this setting the $S_j(t|x;\gamma)$ are easily determined in closed-form and we omit the details. We determine our grid of time-points $\{\tau_j : j = 1, \ldots, q\}$ so that the resulting intervals contain nearly equal numbers of deaths and there is at least one death from each cause in every interval. This is the approach for time-point selection suggested in Larson and Dinse (1985). Our fitted model uses 5 subintervals since the imputed values $\hat{\mathfrak{z}}_j$ from this model were similar to those models with 6 or more subintervals. In addition, likelihood ratio model tests indicate that 5 subintervals yield the best fitting model. The imputed number of osteosarcoma deaths for the censored death times are given in the last column of Table 4. In parentheses, we display these values as a percentage of the number of censored observations. We see that as the mean dose level decreases fewer osteosarcoma deaths are expected. Table 5 shows $LD10$, $LD50$, and $LD90$

estimates, and their respective 95% CIs, calculated using the saddlepoint and the survival-adjusted saddlepoint methods. Imputation increases the estimated values for $LD10$ and $LD50$. Furthermore it yields narrower CIs and provides a more reasonable lower confidence bound for $LD10$; here negative dose values are not meaningful.

5. CONCLUSIONS

We have developed a novel method of CI construction for $LD100p$ using a generalization of the Fieller theorem for exponential families and saddlepoint approximations. The CIs are formed as the intersection of upper and lower confidence bounds, and as such have very good one-tailed coverage, which is particularly important in applications for which one-sided confidence bounds are appropriate. For instance in pharmacology, where one considers the therapeutic response to a drug, an accurate lower confidence bound for LD99, or perhaps more properly ED99, may be of primary interest since it estimates the minimal dose at which at least 99% of the people taking the drug would benefit. On the other hand in environmental toxicology an accurate upper confidence bound for LD1 might be of interest since it estimates the maximal dose which kills no more than 1% of the population.

Simulation results show that, in terms of one-tailed and two-tailed coverage, the proposed methodology generally outperforms competing confidence intervals obtained from the classical Fieller, likelihood ratio, and score methods. In terms of two-tailed coverage, the proposed method is comparable to the Bartlett-corrected likelihood ratio method, but generally outperforms it in terms of one-tailed coverage. We also develop an extension of our methodology which adjusts for survival in a competing risk setting subject to right-censoring. Here semi-Markov competing risk models are developed and the EM algorithm is used to impute binary responses based upon observed censoring times.

REFERENCES

Abdelbasit, K. M., and Plackett, R. L. (1983). "Experimental Design for Binary Data", *Journal of the American Statistical Association*, 78, 90- 98.

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 131-153.

Brown, T., Cai, T., and DasGupta, A. (2002), "Confidence Intervals for a Binomial Proportion and Asymptotic Expansions," *Annals of Statistics,* 30, 160–201.

Bhattacharya, R. N., and Kong, M. (2006), "Consistency and Asymptotic Normality of the Estimated Effective Doses in Bioassay," *Journal of Statistical Planning and Inference*, 137, 643-658.

Butler, R.W. (2007), *Saddlepoint Approximations with Applications.* Cambridge: Cambridge University Press.

Cox, D. R. (1967), "Fieller's Theorem and a Generalization," *Biometrika*, 54, 567-572.

Davison, A.C., and Wang, S. (2002), "Saddlepoint Approximations as Smoothers," *Biometrika*, 89, 933-938.

Dempster, A., Laird, N. and Rubin, D. (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion)", *Journal of the Royal Statistical Society, Series B*, 39, 1-3 8.

Dette, H., Neumeyer, N., and Pilz, K. F. (2005), "A Note on Nonparametric Estimation of the Effective Dose in Quantal Bioassay," *Journal of the American Statistical Association*,100, 503-510.

Faraggi, D., Izikson, P., and Reiser, B. (2003), "Confidence Intervals for the 50 Per Cent Response Dose. *Statistics in Medicine*, 22, 1977–1988.

Fieller (1954), “Some Problems in Interval Estimation,” *Journal of the Royal Statistical Society, Series B*, 16,175-185.

Finney, D.J. (1971), *Probit Analysis.* Cambridge: Cambridge University Press.

Glasby, C. A. (1987), “Tolerance-distribution-free Analyses of Quantal Dose-response Data,” *Journal of the Royal Statistical Society, Series C*, 36, 251-259.

Gleser, L.J. and Hwang, J.T. (1987), “The Nonexistence of $100\,(1-\alpha)\,\%$ Confidence Sets of Finite Expected Diameter in Errors-in-Variables and Related Models,” *Annals of Statistics*, 15, 1351-1362.

Groer, P.G. (1978), “Dose-response Curves and Competing Risks,” *Proceedings of the National Academy of Sciences of the United States of America,* 75, 4087–4091 (correction: 76, 1524).

Harris, P., Hann, M., Kirby, S. P. J., and Dearden, J. C. (1999), “Interval Estimation of the Median Effective Dose for a Logistic Dose-response Curve,” *Journal of Applied Statistics*, 26, 715-722.

Hewlett, P. S. (1947). “The Toxicities of Three Petroleum Oils to Grain Weevils”, *Annals of Applied Biology*, 34, 575 - 585.

Hewlett, P. S. (1975). “Lethal Action of a Refined Mineral Oil on Adult Sitophilus Granarius (L.)”, *Journal of Stored Products Research*, 11, pp. 119–120.

Huang, Y., Kirby, S.P.J., Harris, P., and Dearden, J.C. (2000). Interval Estimation of the 90% Effective Dose: A Comparison of Bootstrap Resampling Methods with Some Large-sample Approaches. Journal of Applied Statistics 27, 63-73 (correction, v. 28, p. 516).

Huang, Y. (2002), “On Large-sample Parametric Approaches for Interval Estimation of the ED90,” *Computational Statistics and Data Analysis*, 40, 527-537.

Hwang, J.T. (1995),.“Fieller’s Problem and Resampling Techniques,” *Statistica. Sinica*, 5, 161-171.

Jee, W.S.S. (1977), “Radiobiology Laboratory Annual Report”, No. COO-119-252, University of Utah, Salt Lake City, Utah, 26-45.

Larson, M.G and Dinse, G.E (1985), “A Mixture Model for the Regression Analysis of Competing Risks Data”, *Journal of the Royal Statistical Society. Series C*, 34, 201-211.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses.* (2nd edition) New York: Wiley.

McCullagh, P. and Nelder, N.A. (1989),.*Generalized Linear Models* (2nd edition), Chapman and Hall, London

Mueller, H.-G., and Wang, J.-L. (1990), “Bootstrap Confidence Intervals for Effective Doses in the Probit Model for Dose-response Data,” *Biometrical Journal*, 32, 529-544.

Pierce, D.A., and Peters, D. (1992), “Practical Use of Higher Order Asymptotics for Multiparameter Exponential Families,” *Journal of the Royal Statistical Society, Series B*, 54, 701-737.

Proschan, F. and Kim, J. (1989), “Piecewise Exponential Estimator of the Survival Function”, *IEEE Transaction on Reliability*, 40, 134-139.

Routledge, R.D. (1994), “Practicing Safe Statistics with the Mid-p,” *Canadian Journal of Statistics*, 22, 103-110.

Schmoyer, R.L., (1984), Sigmoidally Constrained Maximum Likelihood Estimation in Quantal Bioassay,” *Journal of the American Statistical Association* 79, 448–453

Sitter, R.R., and Wu, C.F.J. (1993), “On the Accuracy of Fieller Intervals for Binary Response Data,” *Journal of the American Statistical Association,* 88, 1021-1025.

Skovgaard, I. M. (1987), “Saddlepoint Expansions for Conditional Distributions,” *Journal of Applied Probability*, 24, 875-887.

Stukel, T. A. (2000), “A General Model for Estimating ED100p for Binary Response Dose-response Data,” *American Statistician*, 44, 19-22.

Williams, D.A. (1986),“Interval Estimation of the Median Lethal Dose,” *Biometrics,* 42, 641-645.

TABLES

$\log_{10}(\text{dose})$	Number of subjects	Number affected
x_i	n_i	y_i
.2810	47	47
.2304	50	50
.1523	50	50
.0864	50	46
-.0363	50	25
-.0809	50	0
-.1487	50	2
-.2147	50	1
-.3098	50	0

Table 1. Hewlett Data - Toxicity of *Sitophilus Granarius* (Grain Weevil) to oil

	LD10		LD50		LD90	
	-.0951		-.0173		.0605	
Method	LCL	UCL	LCL	UCL	LCL	UCL
Saddlepoint	-.1182	-.0776	-.0318	-.0004	.0362	.0912
Fieller	-.1190	-.0777	-.0322	-.0000	.0374	.0949
Likelihood Ratio (LR)	-.1168	-.0768	-.0322	-.0009	.0350	.0897
Hwang's Bootstrap	-.1168	-.0772	-.0328	.0000	.0358	.0908
Score	-.1187	-.0778	-.0320	-.0004	.0377	.0938
Bartlett-corrected LR	-.1171	-.0765	-.0323	-.0007	.0345	.0901

Table 2. Hewlett data two-sided 95% CIs for $LD10$, $LD50$, and $LD90$, based on Saddlepoint, Fieller, Likelihood Ratio (LR), Hwang's Bootstrap, Score and Bartlett-corrected LR (BLR) methods.

		n				
β_1	λ	7	10	20	30	50
21	.3	0.812	0.874	0.936	0.945	0.949
21	.2	0.783	0.846	0.925	0.943	0.949
21	.1	0.861	0.910	0.935	0.941	0.946
21	.0	0.850	0.905	0.948	0.952	0.951
14	.3	0.896	0.926	0.945	0.946	0.948
14	.2	0.881	0.916	0.944	0.947	0.948

Table 3. $LD90$ 95% CI coverage rates (%) for Hwang's Bootstrap method for selected values of β_1 and λ.

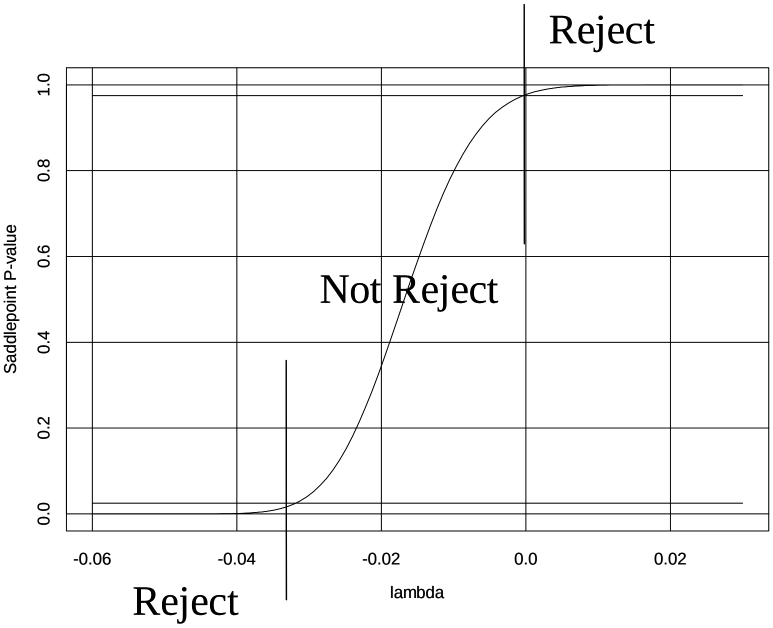
Mean Dose	No.	No.	Num. with	Num.	Imputed Num.
(Ci/kg)	Dogs	Deaths	Osteosarcoma	Censored	Osteoarcoma
2.88	9	9	7	0	n/a
.909	12	12	12	0	n/a
.296	12	12	12	0	n/a
.0951	12	12	10	0	n/a
.0477	14	14	9	0	n/a
.0156	26	14	4	12	1.74 (14.5%)
.0103	38	1	0	37	4.60 (12.4%)
.0054	24	4	0	20	1.97 (7.5%)
.0019	10	3	0	7	0.52 (7.5%)
.0007	13	5	0	8	0.60 (7.5%)
.0000	29	14	0	15	1.11 (7.4%)

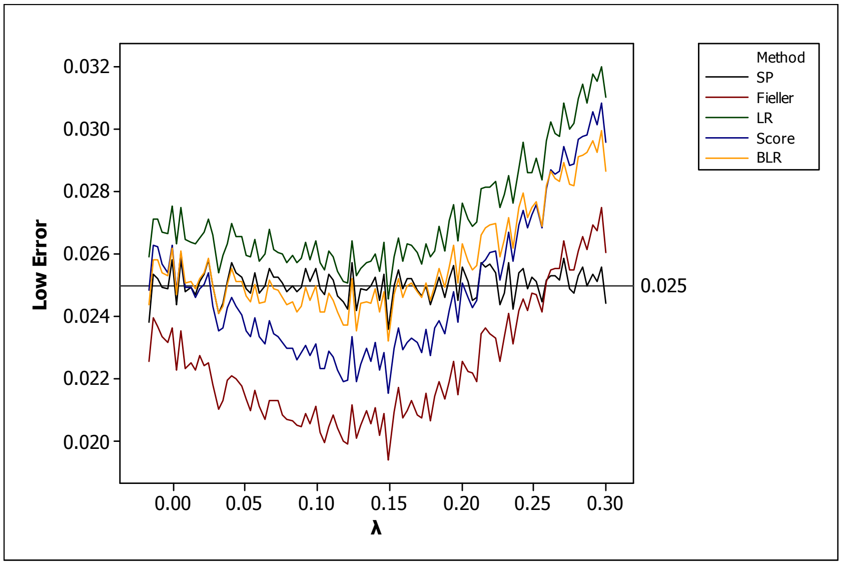
Table 4. Summary of the Beagles data with the imputed number of osteosarcoma deaths for the censored death times.

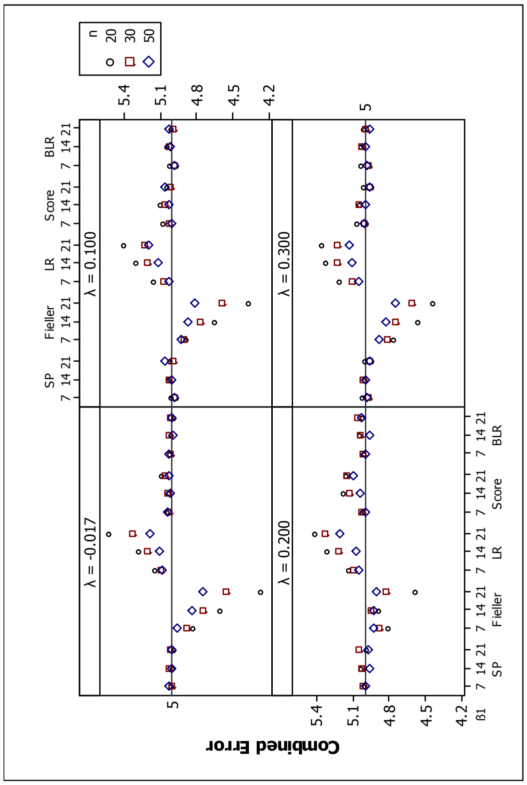
	LD10			LD50			LD90		
Method	MLE	LCL	UCL	MLE	LCL	UCL	MLE	LCL	UCL
Saddlepoint	0.005	-0.023	0.020	0.046	0.034	0.064	0.088	0.067	0.133
Surv-Adj Saddlepoint	0.006	-0.008	0.015	0.047	0.036	0.064	0.088	0.067	0.126

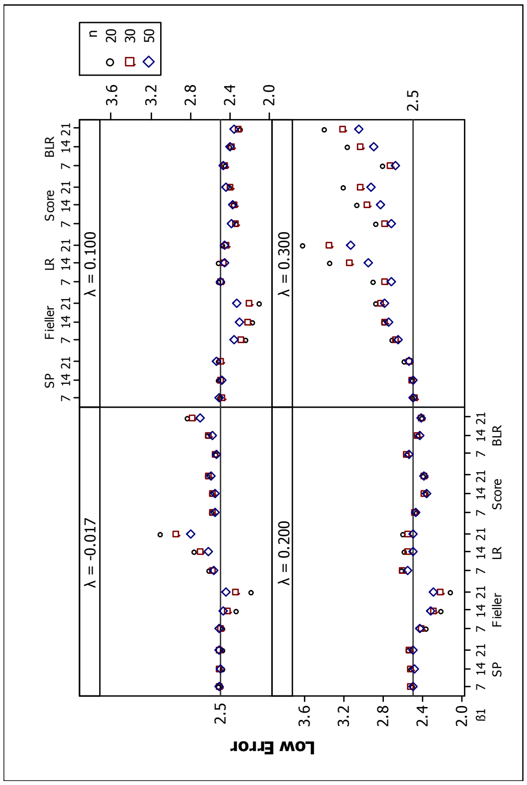
Table 5. Beagle data two-sided 95% CIs for $LD10$, $LD50$, and $LD90$, based on the Saddlepoint and the Survival-adjusted Saddlepoint methods.

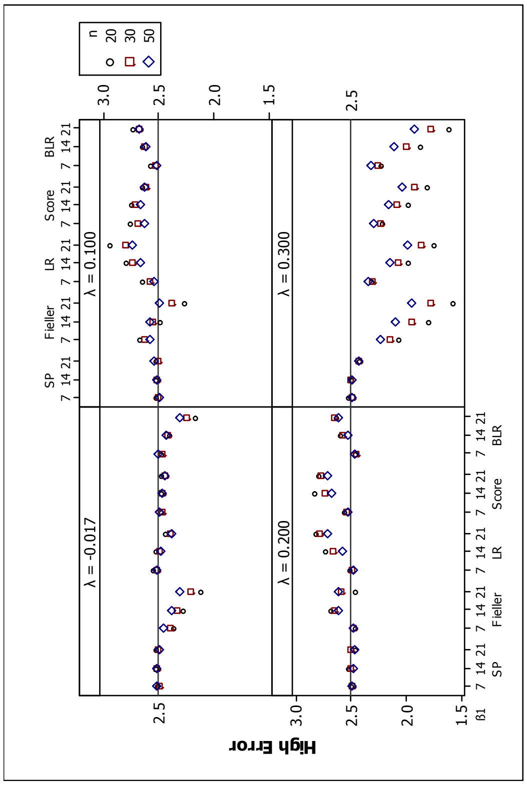
FIGURE TITLES AND LEGENDS

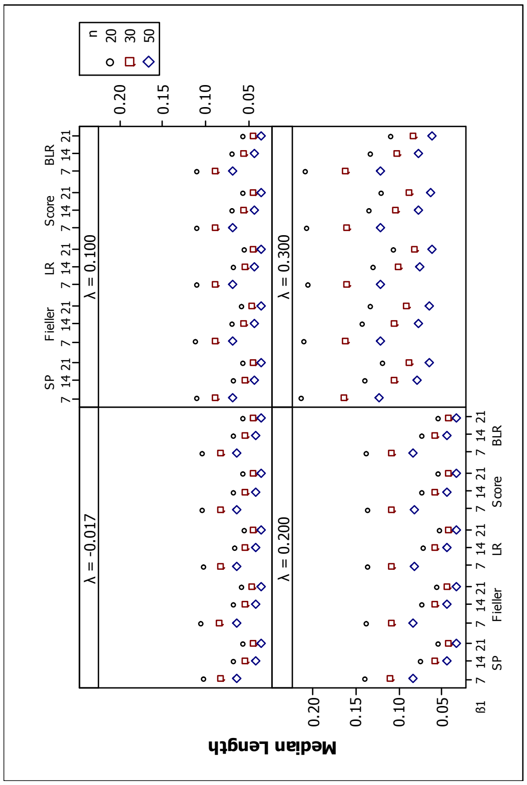

Reject
Not Reject
Reject
Saddlepoint P-value
0.0
0.2
0.4
0.6
0.8
1.0
-0.06
-0.04
-0.02
0.0
0.02
lambda

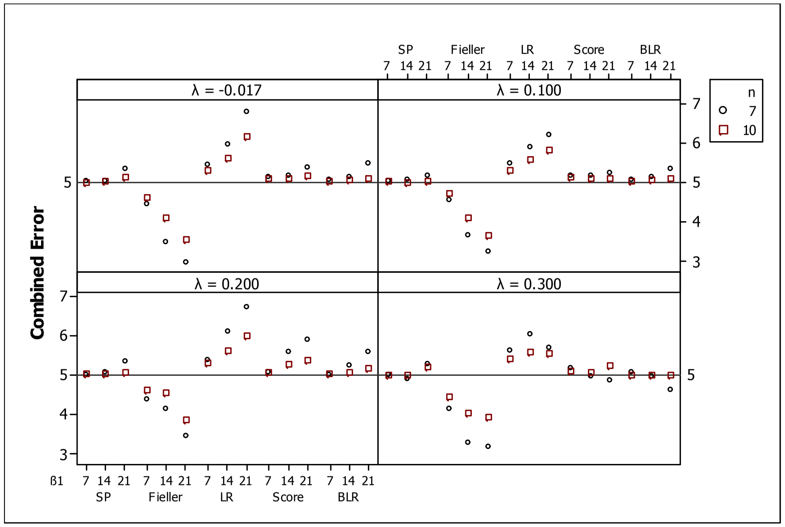

Low Error
0.032
0.030
0.028
0.026
0.024
0.022
0.020
0.00
0.05
0.10
0.15
0.20
0.25
0.30
λ
0.025
Method
SP
Fieller
LR
Score
BLR

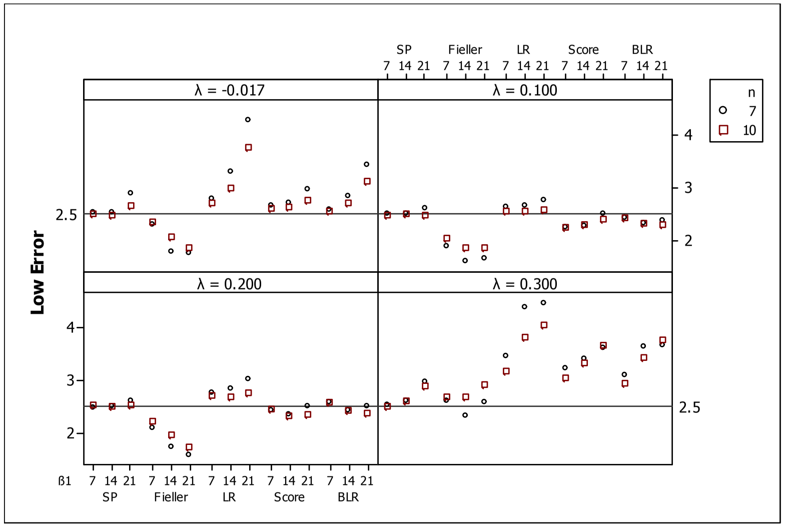

Combined Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
5
5.4
5.1
4.8
4.5
4.2
n
20
30
50

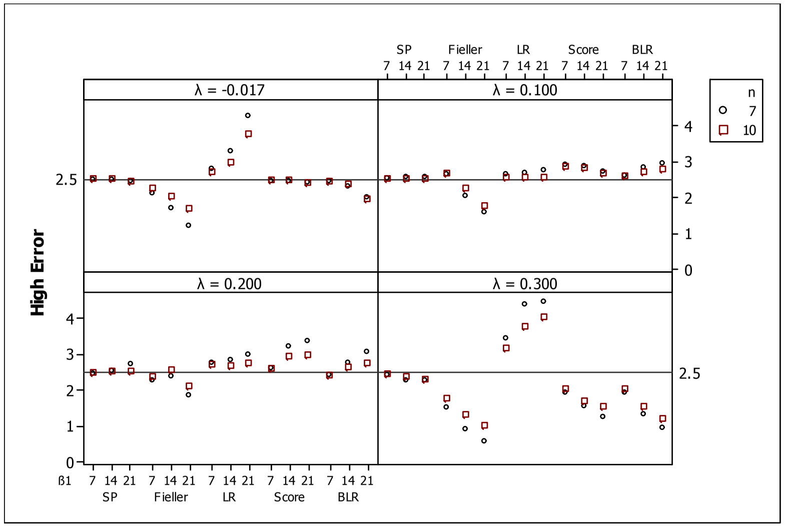

Low Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
2.5
3.6
3.2
2.8
2.4
2.0
n
20
30
50

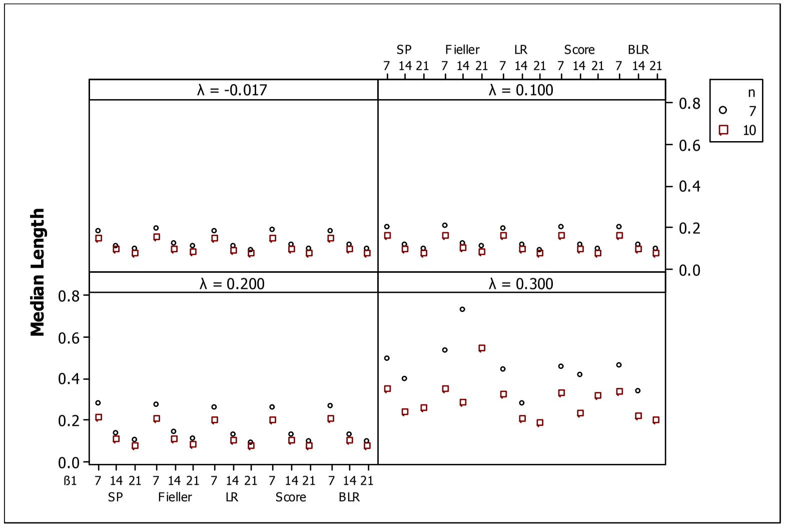

High Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
2.5
3.0
2.0
1.5
n
20
30
50

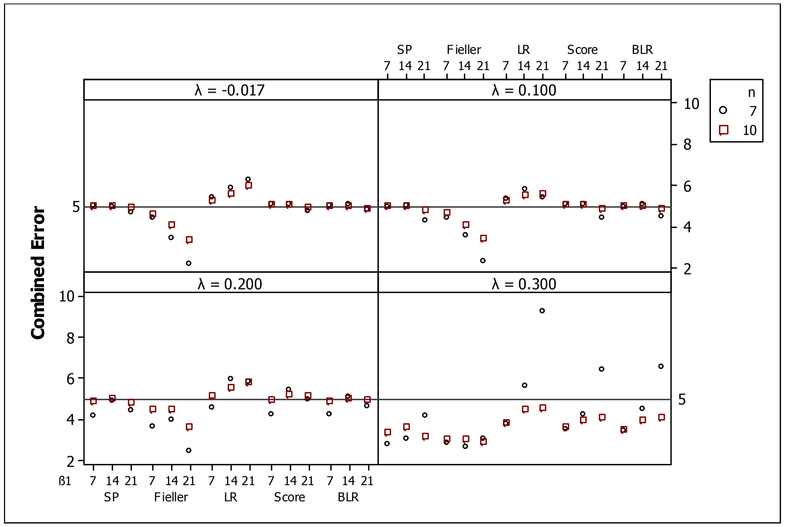

Median Length
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
0.20
0.15
0.10
0.05
n
20
30
50

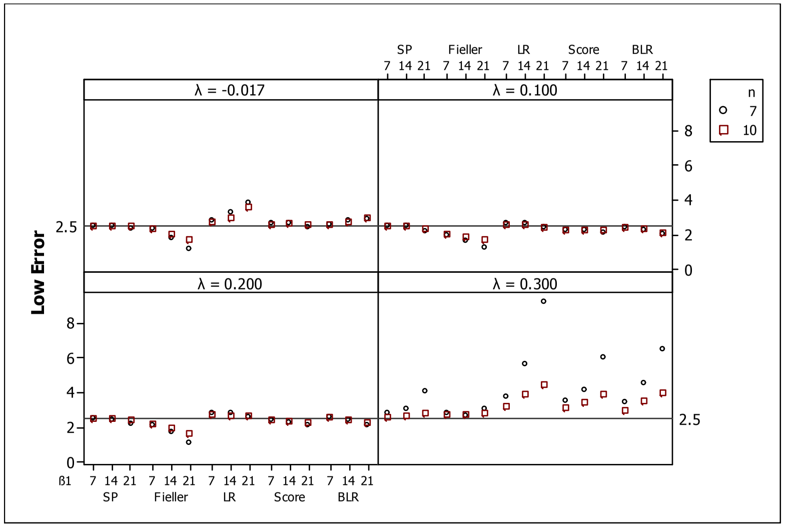

Combined Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
n
7
10
ß1
7 14 21
SP
Fieller
LR
Score
BLR
3
4
5
6
7

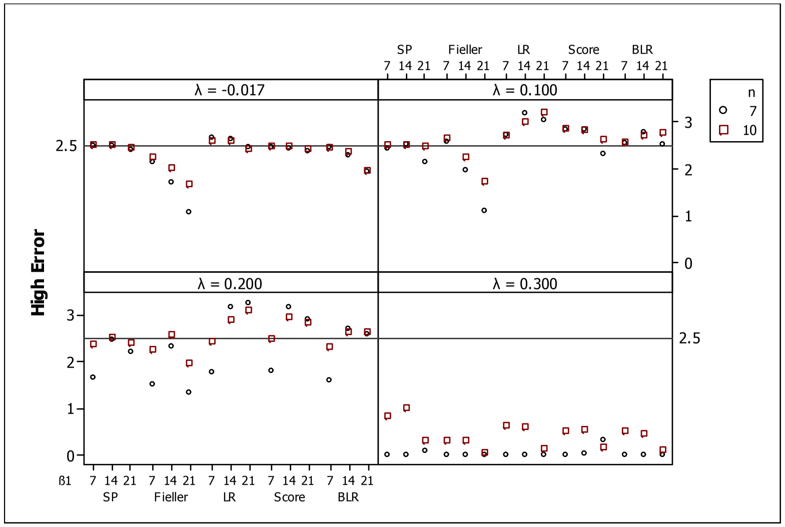

Low Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
n
7
10
ß1
7 14 21
SP
Fieller
LR
Score
BLR
2
2.5
3
4

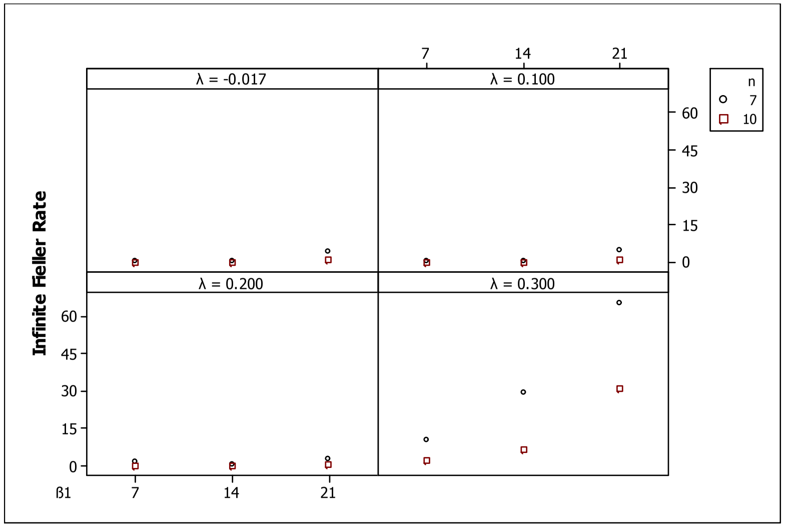

High Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
n
7
10
ß1
7 14 21
SP
Fieller
LR
Score
BLR
0
1
2
2.5
3
4

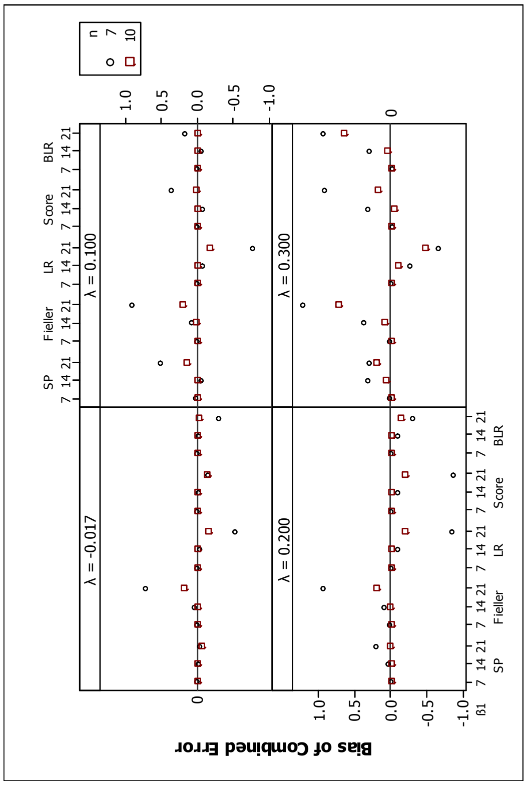

Median Length
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
n
7
10
ß1
7 14 21
SP
Fieller
LR
Score
BLR
0.0
0.2
0.4
0.6
0.8

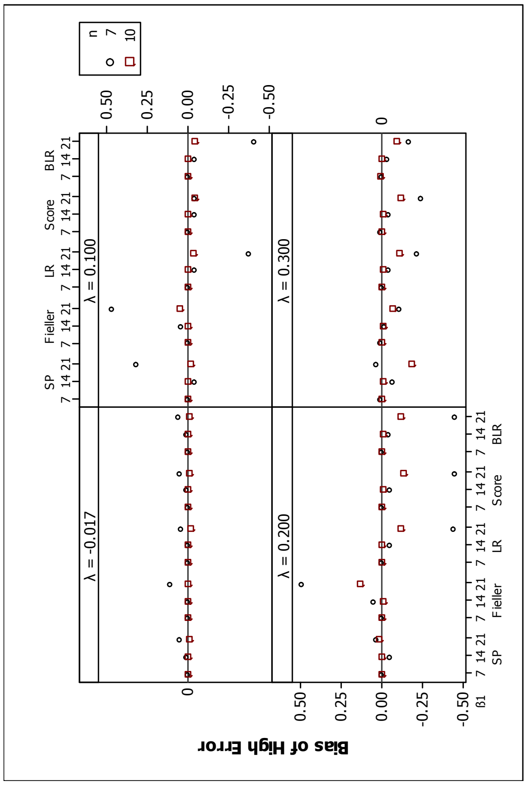

Combined Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
n
7
10
ß1
7 14 21 7 14 21 7 14 21 7 14 21 7 14 21
SP Fieller LR Score BLR
2 4 6 8 10
5

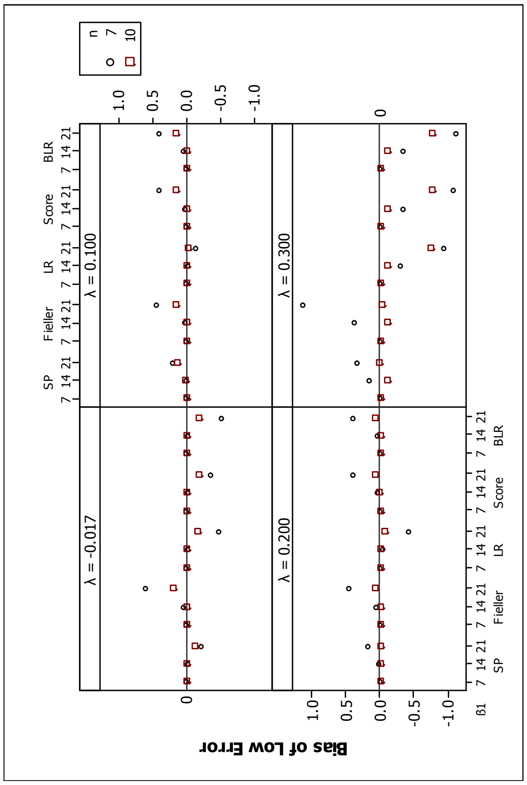

Low Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
n
7
10
ß1
7 14 21 7 14 21 7 14 21 7 14 21 7 14 21
SP Fieller LR Score BLR
0 2 4 6 8
2.5

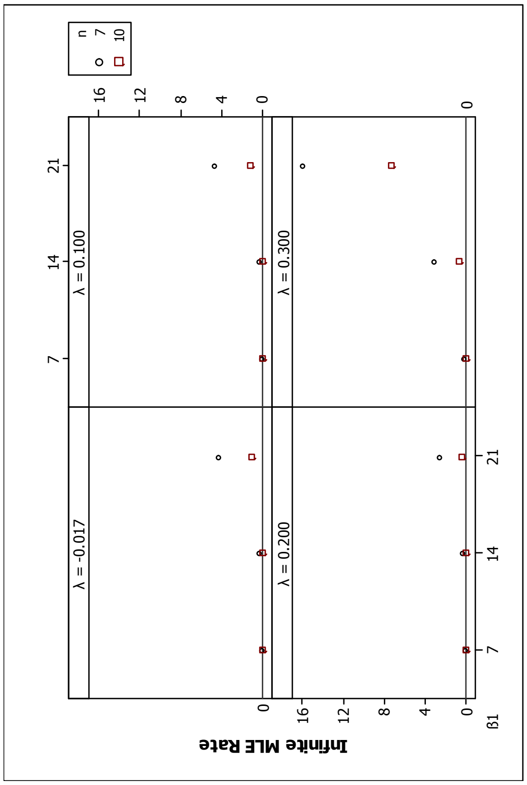

High Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
n
7
10
ß1
7 14 21 7 14 21 7 14 21 7 14 21 7 14 21
SP Fieller LR Score BLR
0 1 2 3
2.5

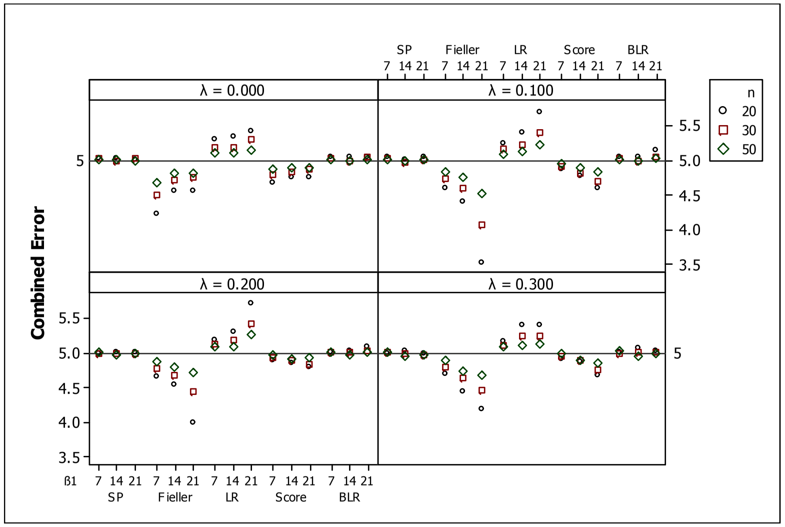

Infinite Fieller Rate
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
n
7
10
ß1
7 14 21
0 15 30 45 60

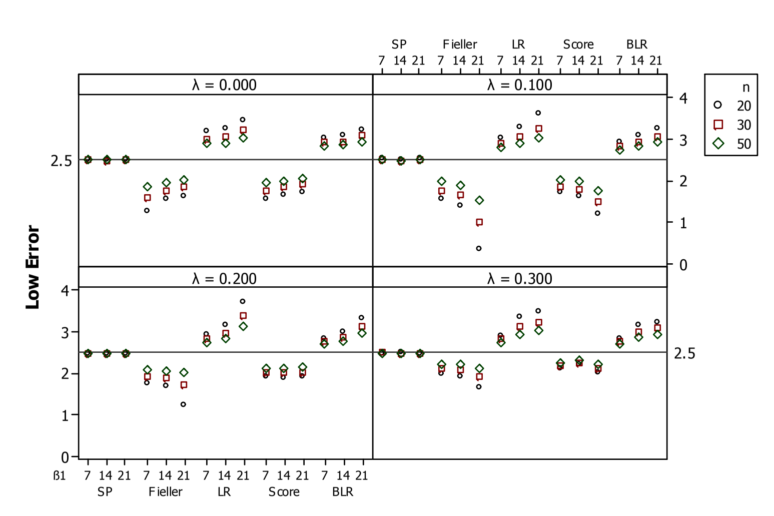

Bias of Combined Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
1.0
0.5
0.0
-0.5
-1.0
0
n
○ 7
□ 10

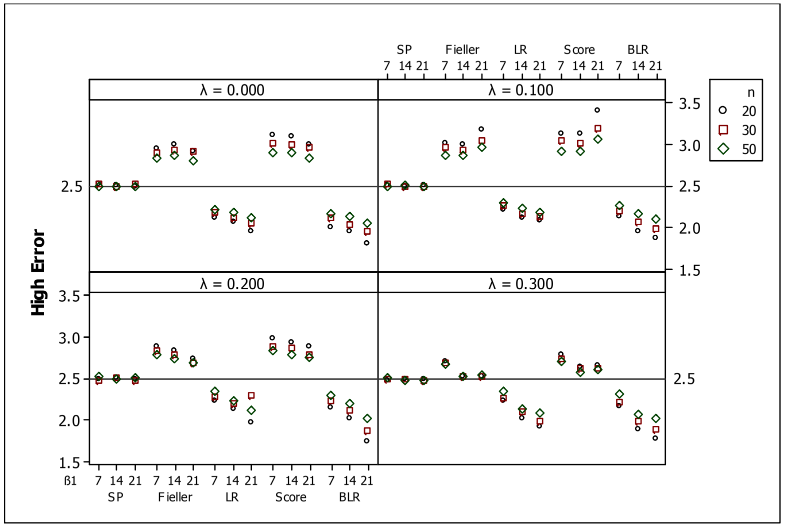

Bias of High Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
0.50
0.25
0.00
-0.25
-0.50
0
n
○ 7
□ 10

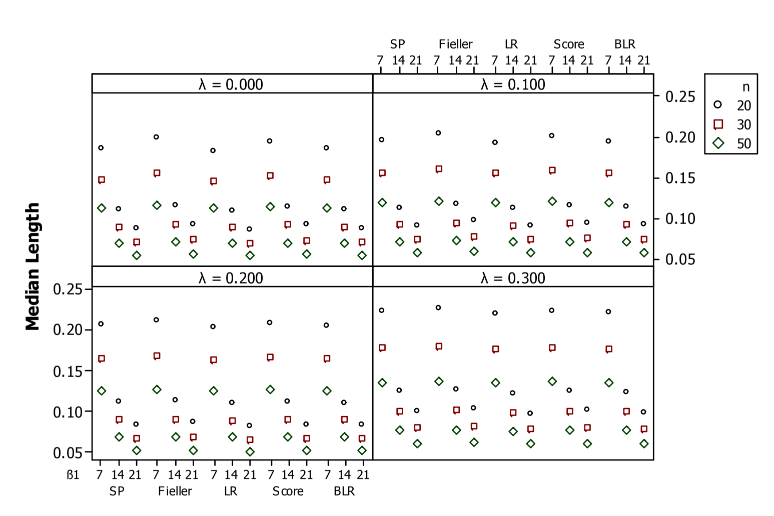

Bias of Low Error
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
1.0
0.5
0.0
-0.5
-1.0
0
n
○ 7
□ 10

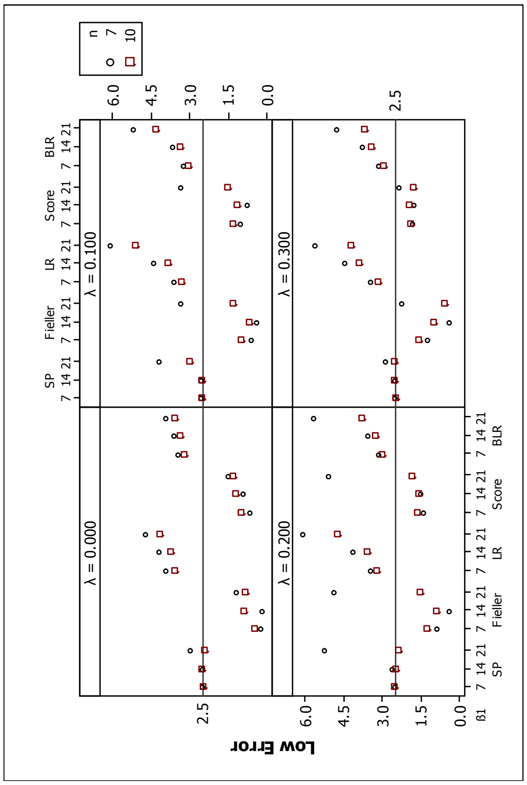

Infinite MLE Rate
λ = -0.017
λ = 0.100
λ = 0.200
λ = 0.300
7
14
21
ß1
16
12
8
4
0
n
○ 7
□ 10

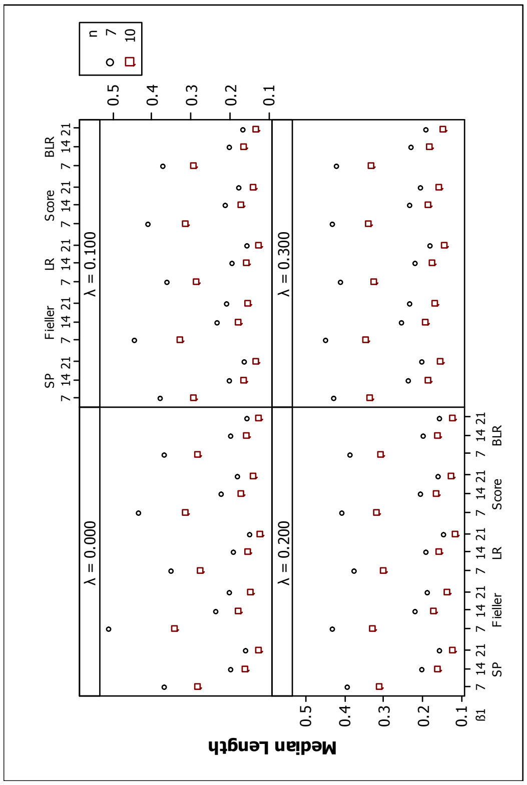

Combined Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
n
20
30
50
5
5.5
5.0
4.5
4.0
3.5
ß1
7 14 21
SP
Fieller
LR
Score
BLR

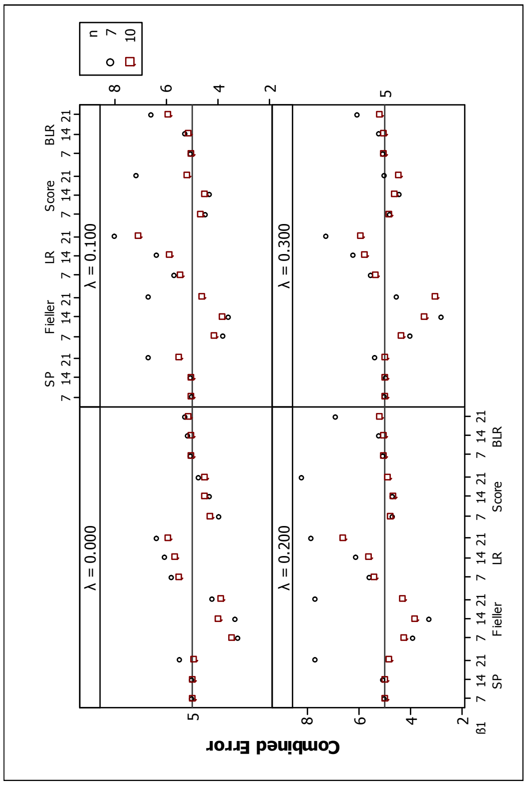

Low Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
n
20
30
50
2.5
4
3
2
1
0
ß1
7 14 21
SP
Fieller
LR
Score
BLR

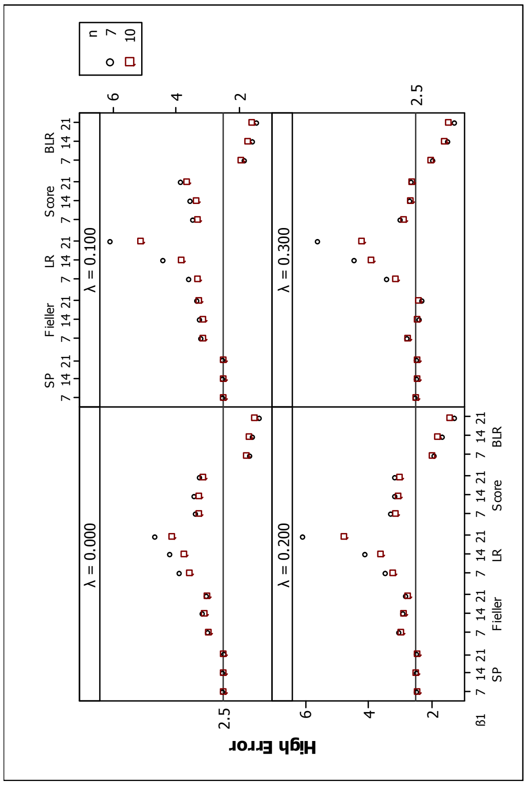

High Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
n
20
30
50
2.5
3.5
3.0
2.0
1.5
ß1
7 14 21
SP
Fieller
LR
Score
BLR

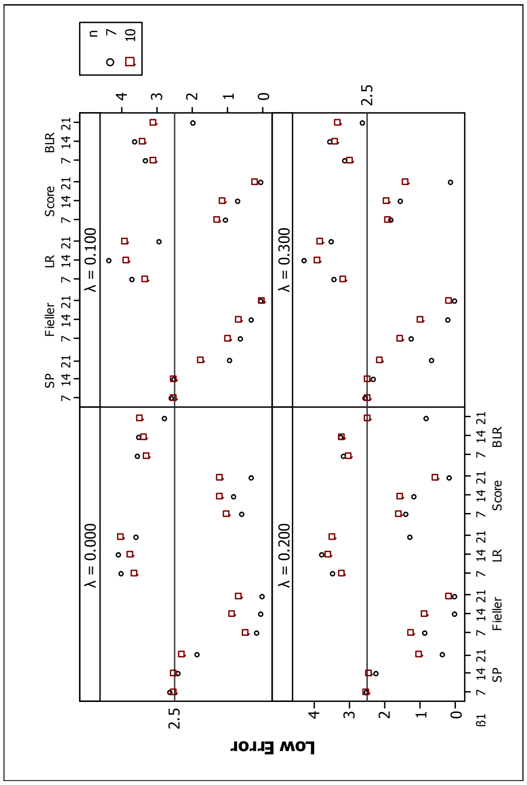

Median Length
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
n
20
30
50
0.25
0.20
0.15
0.10
0.05
ß1
7 14 21
SP
Fieller
LR
Score
BLR

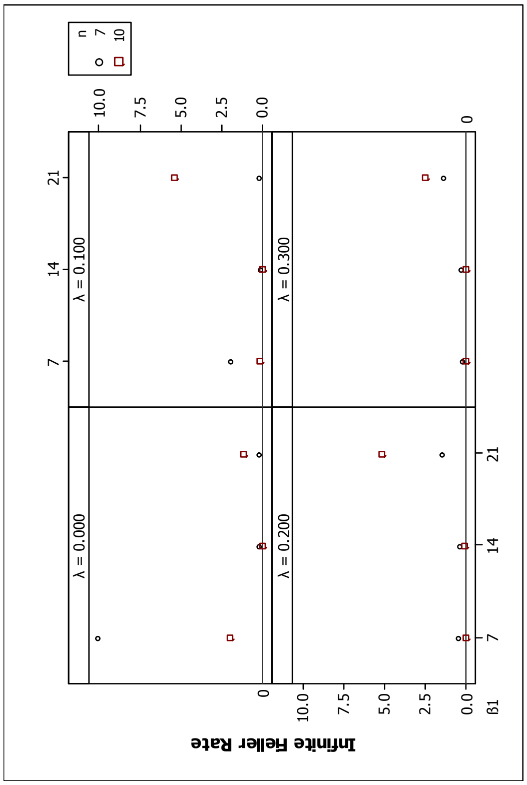

Low Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
β1
0.0
1.5
3.0
4.5
6.0
2.5
n
7
10

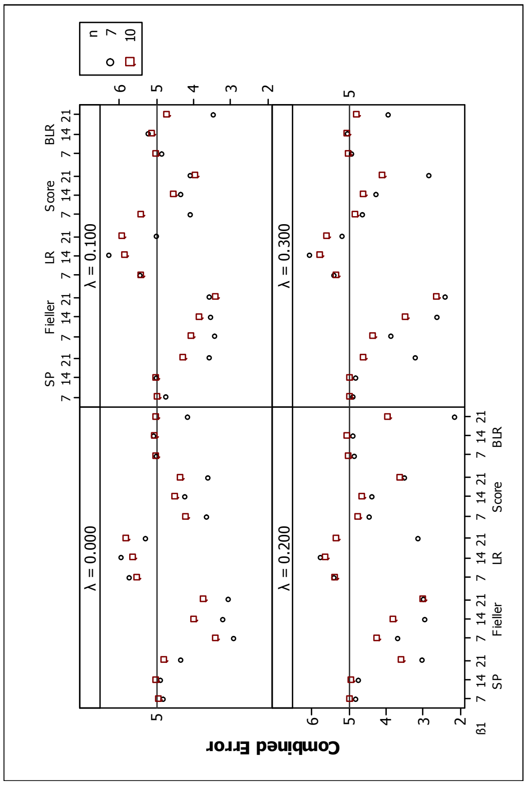

Median Length
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
β1
0.1
0.2
0.3
0.4
0.5
n
7
10

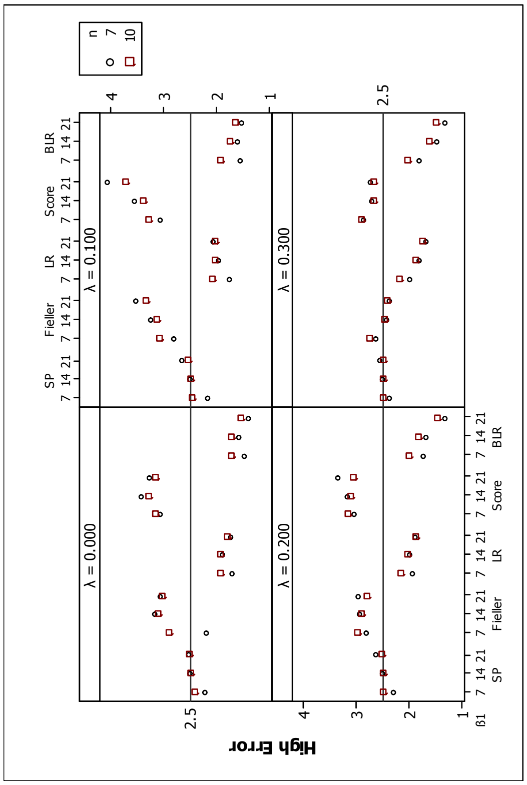

Combined Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
β1
2
4
6
8
5
n
7
10

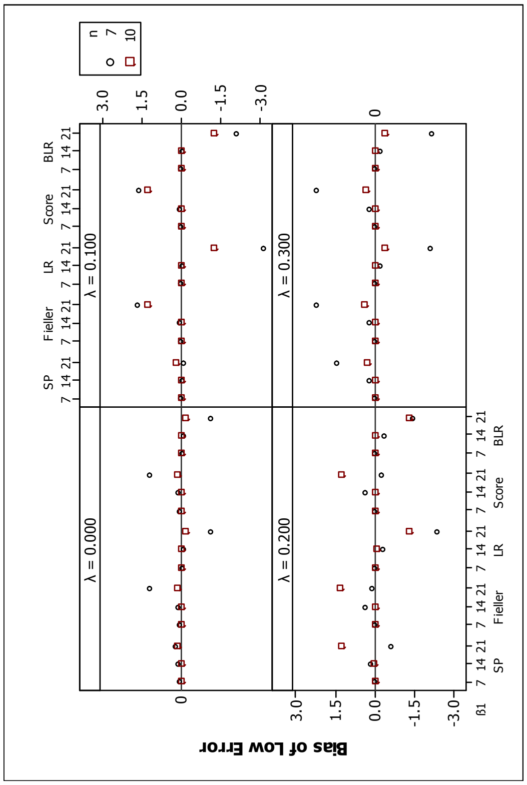

High Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
β1
2
4
6
2.5
n
7
10

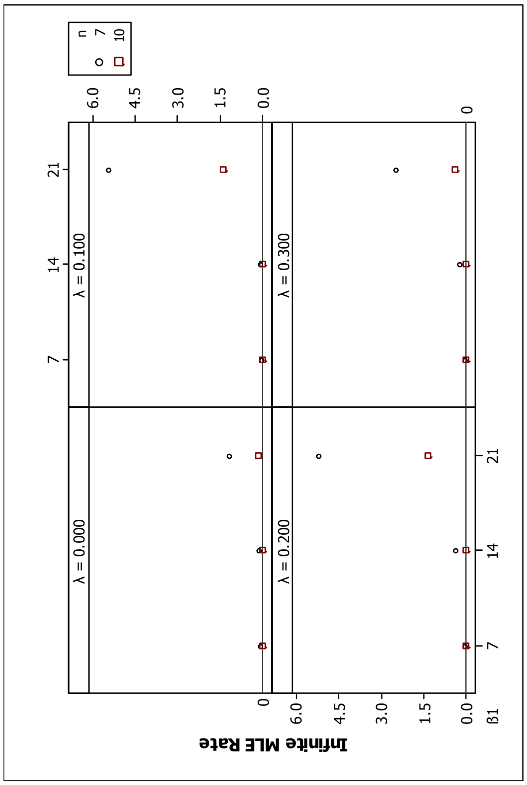

Low Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
2.5
0 1 2 3 4
n
7
10

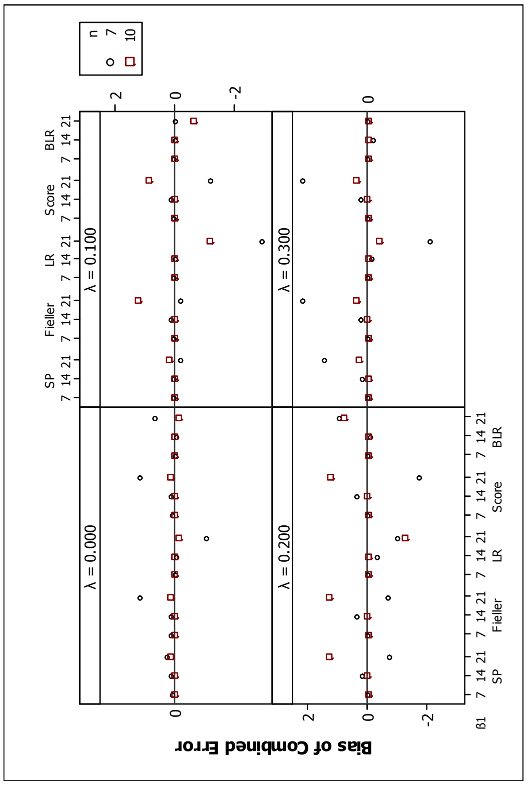

Infinite Fieller Rate
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
7
14
21
ß1
0
0.0 2.5 5.0 7.5 10.0
n
7
10

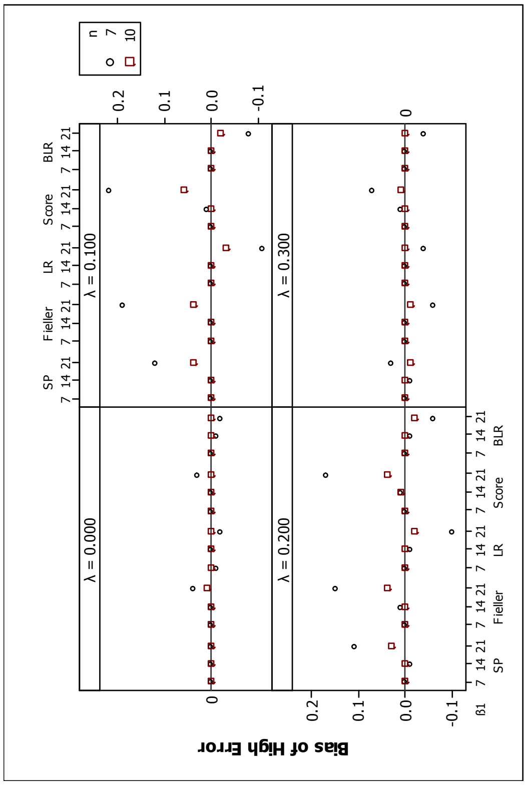

Combined Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
5
2 3 4 5 6
n
7
10


High Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
ß1
2.5
1 2 3 4
n
7
10


Bias of Low Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
β1
3.0
1.5
0.0
-1.5
-3.0
0
n
7
10


Infinite MLE Rate
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
7
14
21
β1
6.0
4.5
3.0
1.5
0.0
0
n
7
10


Bias of Combined Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
β1
2
0
-2
n
7
10


Bias of High Error
λ = 0.000
λ = 0.100
λ = 0.200
λ = 0.300
SP
Fieller
LR
Score
BLR
7 14 21
β1
0.2
0.1
0.0
-0.1
0
n
7
10