Confidence Intervals for Current Status Data

Moulinath Banerjee and Jon A. Wellner

Department of Statistics, University of Michigan, and
Department of Statistics, University of Washington

January 20, 2005

Abstract

The likelihood ratio statistic for testing pointwise hypotheses about the survival time
distribution in the current status model can be inverted to yield confidence intervals. One
advantage of this procedure is that confidence intervals can be formed without estimating the
unknown parameters that figure in the asymptotic distribution of the MLE of the distribution
function. We discuss the likelihood ratio based confidence intervals for the distribution function
and the quantile function and compare these intervals to several different intervals based on the
MLE. The quantiles of the limiting distribution of the MLE are estimated using various methods
including parametric fitting, kernel smoothing, and subsampling techniques. Comparisons are
carried out both for simulated data and on a data set involving time to immunization against
rubella. The comparisons indicate that the likelihood ratio based intervals are preferable from
several perspectives.

Keywords: asymptotic distribution, bootstrap, confidence interval, current status data, kernel
smoothing, quantile estimation, rubella data, subsampling

Running Head: Confidence sets for current status data
1 Introduction

In recent years there has been considerable research on the analysis of interval censored data. Interval censoring happens when the variable of interest is not observed directly but is only known to lie in an interval (one dimensional or multidimensional, as the case may be) in its domain. Such data arise extensively in epidemiological studies and clinical trials, especially in large scale panel studies where the event of interest, which is typically an infection with a disease or some other failure (like organ failure), is not observed exactly but is only known to happen between two consecutive examination times. In particular, large scale HIV/AIDS studies typically yield various types of interval censored data where interest centers on the distribution of time to HIV infection, but the exact time of infection is only known to lie between two consecutive follow-ups at the clinic. In this paper, we are interested in estimating the failure time distribution in the most basic version of the interval censoring model, known as the current status model or Case 1 interval censoring. Here, the individual is checked only at a single point in time and the status of the individual ascertained: 1 if the infection/failure has occurred by the time they are checked and 0 otherwise. We introduce a new likelihood ratio based method for interval estimation of the distribution of time to event in the current status model and compare it with several existing methods.

We now formally introduce the current status model.

The Current Status Censoring Model: Let \((X_1, T_1), (X_2, T_2), \ldots, (X_n, T_n)\) be \(n\) i.i.d. pairs of random variables. For each \(i\), \(X_i\) is distributed as \(F\), \(T_i\) is distributed as \(G\), and \(X_i\) is independent of \(T_i\). The distributions \(F\) and \(G\) are continuous and concentrated on the positive half-line. More concretely, we can think of \(n\) individuals with \(X_i\) being the failure/survival time and \(T_i\) the “observation time” for the \(i\)th individual, respectively. For the \(i\)th individual we observe the vector \((\Delta_i, T_i)\), where \(\Delta_i = 1\{X_i \leq T_i\}\) is the indicator of a failure before \(T_i\). We are interested in estimating \(F\), based on the current status data, and more specifically, in estimating \(F(t_0)\), the value of \(F\) at \(t_0\), a fixed time point. We are also interested in making inference on the quantiles of \(F\); i.e., estimate \(F^{-1}(\theta_0)\) for some \(0 < \theta_0 < 1\).

The current status model introduced above is in some sense a fundamental model and is a natural starting point for a large number of censored data models used in practice. It is very different from right-censored models in that one does not observe the actual failure time itself; consequently the
estimation of the survival distribution $F$ is harder in this situation. This is reflected in a slower rate of convergence of the nonparametric maximum likelihood estimator for $F$ ($n^{1/3}$) as we will see shortly, as opposed to the usual $\sqrt{n}$ rate that one encounters with right-censored models. The current status model easily generalizes to the Case $k$ interval censoring model, where an individual with survival time $X$ is observed at $k$ time-points $T_0 \equiv 0 < T_1 < T_2 < \ldots < T_k < T_{k+1} \equiv \infty$, with the $k$ observation times being random, and one observes in which interval, $(T_i, T_{i+1}]$, the individual fails. The problem is then, as in the current status example, to estimate $F$, the distribution function of $X$. See, for example Groeneboom & Wellner (1992) and Groeneboom (1996). The Case $k$ interval censoring model can be generalized further to mixed case interval censoring, where $k$ itself is a random variable, to allow greater flexibility in modeling (see for example, Schick & Yu (2000)). The Case $k$ or mixed case interval censoring models generalize naturally to counting process models where we have a counting process $N(t)$ associated with each individual and only counts at the observation times are recorded. Based on these, one seeks to estimate $E(N(t)) \equiv \Lambda(t)$, the mean function of the counting process. Such models are important when one deals with recurrent events – for example, a series of attacks or seizures as a patient is being monitored over time. For more discussion on these issues, see for example, Wellner & Zhang (2000).

While this paper will deal exclusively with (different estimation procedures for) the Case 1 interval censoring model, there is strong evidence to suggest that similar approaches can be taken with the more general censoring models considered above.

The rest of this paper is organized as follows. Section 2 describes several different procedures for the construction of pointwise confidence intervals for $F$ or $F^{-1}$. A natural way of constructing such intervals is to use the asymptotic distribution of the nonparametric MLE of $F$ or $F^{-1}$; this requires estimating the quantiles of the limit distribution, which can be done in several different ways (using resampling techniques, smoothing or parametric fits). A new method for construction of confidence sets that proceeds via inversion of the likelihood ratio test for testing pointwise hypotheses about $F$ or $F^{-1}$ is introduced. A major advantage of the likelihood ratio method is the fact that the asymptotic distribution of the likelihood ratio statistic is free of nuisance parameters; consequently, the computation of asymptotic critical values does not require estimation of nuisance parameters in the underlying model. Section 3 discusses the issues involved with these procedures and assesses the relative merits of these methods through simulation studies. Our simulations indicate several advantages of the likelihood ratio method over their MLE based counterparts. In Section 4, the different methods are applied to a data-set involving time to immunization by
Rubella in a population of Austrian males. Section 5, the appendix, contains proofs of some of the results in the previous sections.

We end this section with some notation. For positive constants $a$ and $b$ and a two-sided Brownian motion $W(h)$, we denote the process $aW(h) + bh^2$, where $h$ varies over the reals, by $X_{a,b}(h)$. We denote the slope (right derivative) of the greatest convex minorant (GCM) of $X_{a,b}$ by $g_{a,b}$. Thus, $g_{1,1}$ is the slope process of the GCM of the process $W(h) + h^2$ on the line. We denote by $Z_{a,b}$, the almost surely unique minimizer of the process $X_{1,1}(h)$ on the line. For details see Banerjee & Wellner (2001) or Banerjee (2000).

2 Estimation procedures for $F$ and $F^{-1}$

We denote the distribution of $(\Delta, T)$ under $(F, G)$ by $P_{F,G}$. The log-likelihood based on $n$ i.i.d. observations $\{(\Delta_1, T_1), (\Delta_2, T_2), \ldots, (\Delta_n, T_n)\}$ is then given by

$$
\log L_n(F) = \sum_{i=1}^{n} (\Delta_i \log F(T_i) + (1 - \Delta_i) \log(1 - F(T_i)))
= n\mathbb{P}_n(\Delta \log F(T) + (1 - \Delta) \log(1 - F(T)))
(1)
$$

where $\mathbb{P}_n$ is the empirical measure of the observations $\{(\Delta_i, T_i)_{i=1}^{n}\}$. The methods in this paper are based on maximization of the likelihood function (both without and under constraints on $F$) and are described below. To ensure that the procedures used in this paper are correct, we need to assume that both $F$ and $G$ are continuously differentiable in a neighborhood of $t_0$, the point of interest, with positive derivatives $f$ and $g$.

2.1 Pointwise confidence sets for the distribution function

The MLE based method: This is based on the asymptotic distribution of the MLE of $F(t_0)$ in the interval-censoring model. The MLE of $F(t_0)$ is $\hat{F}_n(t_0)$ where $\hat{F}_n$ is the NPMLE of $F$ based on the current status data. From Groeneboom & Wellner (1992), Theorem 5.1, page 89. it follows that

$$
n^{1/3} (\hat{F}_n(t_0) - F(t_0)) \rightarrow_d \left(\frac{4f(t_0)F(t_0)(1 - F(t_0))}{g(t_0)}\right)^{1/3} Z \equiv CZ
$$

where $C = C(F, f, g, t_0)$. A 95% confidence interval for $F(t_0)$ is then given by

$$
[\hat{F}_n(t_0) - n^{-1/3} \hat{Q}_{.975}, \hat{F}_n(t_0) - n^{-1/3} \hat{Q}_{.975}],
$$
where \( Q_{.975} \) is a consistent estimator of \( Q_{.975} \), the 97.5'th percentile of the limiting random variable \( CZ \). But \( Q_{.975} \) is simply \( C \times .99818 \) where .99818 is the 97.5'th percentile of \( Z \); see Groeneboom & Wellner (2001), where quantiles of \( Z \) are computed. Since \( C \) involves the unknown parameters \( F(t_0) \), \( g(t_0) \), and \( f(t_0) \), we estimate \( C \) by

\[
\hat{C}_n = \left( \frac{4 \hat{f}_n(t_0) \hat{F}_n(t_0) (1 - \hat{F}_n(t_0))}{\hat{g}_n(t_0)} \right)^{1/3},
\]

where \( \hat{f}_n \) and \( \hat{g}_n \) are estimates of \( f \) and \( g \). An asymptotic 95% confidence interval is then given by

\[
\left[ \hat{F}_n(t_0) - n^{-1/3} \hat{C}_n \times .99818, \hat{F}_n(t_0) + n^{-1/3} \hat{C}_n \times .99818 \right].
\]

In this paper estimates of \( f \) and \( g \) are obtained by kernel smoothing with bandwidths chosen using cross–validation techniques (to be elaborated later). Parametric estimation of \( f \) and \( g \) are considered in the context of the Rubella data analysis.

**Parametric estimation of the nuisance parameters**: Another possible option might be to estimate \( f(t_0) \) and \( g(t_0) \) based on parametric models for \( f \) and \( g \). This approach was used by Keiding et al. (1996) with Weibull models for both \( f \) and \( g \). For more details see Section 4.

**Subsampling based methods**: The subsampling technique followed here is from Politis, Romano & Wolf (1999) and is part of a general theory for obtaining confidence regions. The basic idea is to approximate the sampling distribution of a statistic, based on the values of the statistic computed over smaller subsets of the data. In the context of interval censored data, we have i.i.d. observations \( U_1, U_2, \ldots, U_n \) from the model (with \( U_i \equiv (\Delta_i, T_i) \)) and \( \hat{F}_n(t_0) = \hat{\theta}_n \) is based on this i.i.d. sample. We also know that \( n^{1/3} (\hat{F}_n(t_0) - F(t_0)) \) has a limit distribution \( J \). To obtain large sample confidence regions for \( \theta_0 \equiv F(t_0) \), we consider \( Y_1, Y_2, \ldots, Y_{N_n} \), where the \( Y_i \)'s are the \( N_n \equiv \binom{n}{b} \) subsets of \( \{U_1, U_2, \ldots, U_n\} \) of size \( b \) listed in any order. Let \( \hat{\theta}_{n,b,i} \) be the value of the NPMLE of \( F \) at \( t_0 \) computed at data set \( Y_i \). Now define

\[
L_{n,b}(x) = N_n^{-1} \sum_{i=1}^{N_n} 1 \{ b^{1/3} (\hat{\theta}_{n,b,i} - \hat{\theta}_n) \leq x \}.
\]

Let \( c_{n,b}(\beta) = \inf \{ x : L_{n,b}(x) \geq \beta \} \). Now let \( b \to \infty \) as \( n \to \infty \), but in such a way that \( b/n \to 0 \).

It then follows from Theorem 2.2.1 of Politis, Romano & Wolf (1999), and the fact that \( J \) is continuous, that for any \( 0 < \beta < 1 \),

\[
P_{F,G} \left( n^{1/3} (\hat{\theta}_n - \theta_0) \leq c_{n,b}(\beta) \right) \to \beta.
\]

It follows easily that for any \( 0 < \alpha < 0.5 \),

\[
P_{F,G} \left( (c_{n,b}(\alpha/2) < n^{1/3} (\hat{\theta}_n - \theta_0) \leq c_{n,b}(1 - \alpha/2) \right) \to 1 - \alpha.
\]
Thus an asymptotic level $1 - \alpha$ confidence set for $\theta_0$ is given by
\[
\left[ \hat{\theta}_n - n^{-1/3}c_{n,b}(1 - \alpha/2) , \hat{\theta}_n - n^{-1/3}c_{n,b}(\alpha/2) \right].
\]
This approach can be slightly modified to yield symmetric subsampling based intervals. Instead of considering the limiting distribution of $n^{1/3}(\hat{\theta}_n - \theta_0)$, one considers the limiting distribution of $n^{1/3} | \hat{\theta}_n - \theta_0 |$, say $\tilde{J}$. Let
\[
\tilde{L}_{n,b}(x) = N_n^{-1} \sum_{i=1}^{N_n} 1 \{ b^{1/3} | \hat{\theta}_{n,b,i} - \hat{\theta}_n | \leq x \}
\]
and $\tilde{c}_{n,b}(\beta) = \inf \{ x : \tilde{L}_{n,b}(x) \geq \beta \}$. As before, if $b \to \infty$ as $n \to \infty$, but in such a way that $b/n \to 0$,
\[
P_{F,G} \left( n^{1/3} | \hat{\theta}_n - \theta_0 | \leq \tilde{c}_{n,b}(\beta) \right) \to \beta,
\]
whence it follows that an approximate level $1 - \alpha$ confidence interval for $\theta$ is $[\hat{\theta}_n - n^{-1/3}\tilde{c}_{n,b}(1 - \alpha), \hat{\theta}_n + n^{-1/3}\tilde{c}_{n,b}(1 - \alpha)]$. Note that this is symmetric about $\hat{\theta}_n$. Symmetric subsampling intervals often have nicer properties than their more general counterparts in finite samples. In fact, simulation studies showed this to be the case in the current status model; hence in this paper we have used symmetric subsampling intervals.

Since $N_n = \binom{n}{b}$ can be large, $\tilde{L}_{n,b}$ can be difficult to compute. But one can estimate $\tilde{L}_{n,b}$ via sampling. Choose $\{I_1, I_2, \ldots, I_S\}$ randomly with or without replacement from $\{1, 2, \ldots, N_n\}$, approximate $\tilde{L}_{n,b}(x)$ by
\[
\hat{L}_{n,b}(x) \equiv S^{-1} \sum_{i=1}^{S} 1 \{ b^{1/3} | \hat{\theta}_{n,b,i} - \hat{\theta}_n | \leq x \}
\]
and compute the $\tilde{c}_{n,b}(\beta)$’s based on $\hat{L}_{n,b}$ to get confidence regions. If $S \to \infty$ with $n$, then the confidence regions obtained thus are still asymptotically level $1 - \alpha$.

The likelihood ratio based method: The likelihood ratio statistic for testing the hypothesis $F(t_0) = \theta_0$ is given by
\[
2 \log(\lambda_n) = 2 \left( \log L_n(F_n) - \log L_n(F_n^0) \right),
\]
where $F_n$ is the unconstrained MLE and $F_n^0$ is the constrained MLE under the null hypothesis. It is shown in Banerjee & Wellner (2001), Theorem 2.6, that under the above assumptions on $F$ and $G$,
\[
2 \log(\lambda_n) \to_d \int \left\{ (g_{1,1}(z))^2 - (g_{1,1}^0(z))^2 \right\} dz \equiv D.
\]
Confidence sets of level $1 - \alpha$ with $0 < \alpha < 1$ are obtained by inverting the acceptance region of the likelihood ratio test of size $\alpha$; more precisely if $2 \log \lambda_n(\theta)$ is the likelihood ratio statistic evaluated under the null hypothesis $H_0 : F(t_0) = \theta$, then the set of all $\theta$'s for which $2 \log \lambda_n(\theta)$ is not greater than $d_\alpha$ where $d_\alpha$ is the $1 - \alpha$'th percentile of $D$, gives a limiting level $1 - \alpha$ confidence set for $\theta$. The following proposition guarantees that the confidence sets obtained via inversion of the likelihood ratio statistic as described above have asymptotically correct coverage probability.

Denote the confidence set of (approximate) level $1 - \alpha$ based on a sample of size $n$ from the interval censoring problem by $C_{n,\alpha}$. Thus $C_{n,\alpha} = \{\theta : 2 \log \lambda_n(\theta) \leq d_\alpha\}$. Denote the true distribution function of the event time by $F_0$ and let $\theta_0 = F_0(t_0)$.

**Proposition.** Suppose that $F_0$ and $G$ are continuously differentiable densities $f$ and $g$ in a neighborhood of $t_0$ with $f(t_0), g(t_0) > 0$. Then

$$P_{F_0,G}(\theta_0 \in C_{n,\alpha}) \rightarrow P(D \leq d_\alpha) = 1 - \alpha.$$ 

**Proof.** This follows on noting that

$$P_{F_0,G}(\theta_0 \in C_{n,\alpha}) = P_{F_0,G}(2 \log \lambda_n(\theta_0) \leq d_\alpha) \rightarrow P(D \leq d_\alpha) = 1 - \alpha,$$

by appealing directly to Theorem 2.6 of Banerjee & Wellner (2001).

That the sets $C_{n,\alpha}$ are closed bounded intervals under mild conditions is guaranteed by Theorem 3.9.1 of Banerjee (2000).

### 2.2 Pointwise estimation of quantiles

Given a (one-dimensional) distribution function $H$, we can construct $H^{-1}$, the “inverse distribution function” by setting $H^{-1}(p) = \inf \{x : H(x) \geq p\}$. It is easy to show that $H^{-1}$ is well-defined and that $H(H^{-1}(p) -) \leq p \leq H(H^{-1}(p))$. If $H$ is continuous, then $H(H^{-1}(p) -) = p = H(H^{-1}(p))$. We call $H^{-1}(p)$, the $p$'th quantile of the distribution function $H$. Statisticians often refer to a $p$'th quantile rather than the $p$'th quantile, defining any number $x$ satisfying $H(x-) \leq p \leq H(x)$ to be a $p$'th quantile. We however prefer to have a well-defined notion of the quantile, largely for the purpose of avoiding ambiguity in what follows.

An important problem, in the context of interval censored data, is to estimate the quantiles of the
distribution function of the survival time. With the rubella data set (dealt with in Section 4), for example, we might be interested in estimating the age by which 50% of individuals in the male population are immunized against the disease. With $F$, as before, denoting the distribution of time to immunization, this amounts to estimating $F^{-1}(\theta_0)$ with $\theta_0 = 0.5$. In this section we present results on quantile estimation in the interval-censoring problem; more specifically we deduce the asymptotic distribution of the quantile estimates based on the MLE of $F$, and also deduce the asymptotic distribution of the likelihood ratio statistic for testing a null hypothesis of the form $F^{-1}(\theta_0) = t_0$ for a fixed $\theta_0$. We then use these results to obtain confidence sets for the quantiles of $F$ and as before, compare various methods, by which this can be done.

**Asymptotic distribution of the MLE of $F^{-1}$**: We next obtain the asymptotic distribution of the estimate of $t_0 = F^{-1}(\theta_0)$ based on the MLE $F_n$, of $F$, which is given by $F_n^{-1}(\theta_0)$.

**Lemma 1** If $F$ and $G$ are continuously differentiable in a neighborhood of $t_0 = F^{-1}(\theta_0)$, with positive densities $f(t_0)$ and $g(t_0)$, we have, for $\lambda \in \mathbb{R}$, $x \in \mathbb{R}$,

$$P \left\{ n^{1/3} \left( F_n^{-1}(\theta_0) + \lambda n^{-1/3} - F^{-1}(\theta_0) \right) \leq x \right\} \rightarrow P \left( \frac{4 \theta_0 (1 - \theta_0)}{g(t_0) f(t_0)^2} \right)^{1/3} Z \leq x - \frac{\lambda}{f(t_0)}.$$

The above lemma is proved in the appendix. On setting $\lambda = 0$ in the above proposition we have the following theorem, which gives us the asymptotic distribution of the MLE of the quantiles of the survival distribution function.

**Theorem 1** Consider the current status model. Let $F$ and $G$ be continuously differentiable in a neighborhood of $t_0 = F^{-1}(\theta_0)$, with positive densities $f(t_0)$ and $g(t_0)$. Then

$$n^{1/3} (F_n^{-1}(\theta_0) - F^{-1}(\theta_0)) \rightarrow_d \left( \frac{4 \theta_0 (1 - \theta_0)}{g(t_0) f(t_0)^2} \right)^{1/3} Z = \frac{1}{f(t_0)} \left( \frac{4 f(t_0) \theta_0 (1 - \theta_0)}{g(t_0)} \right)^{1/3} Z.$$

The above result can be used to get an asymptotic 95% confidence interval for $F^{-1}(\theta_0)$; note however that the constants involved in the limit distribution need to be estimated, as before. An approximate asymptotic confidence interval is given by:

$$\left[ F_n^{-1}(\theta_0) - \hat{D}_n n^{-1/3} q_{Z,0.975}, F_n^{-1}(\theta_0) + \hat{D}_n n^{-1/3} q_{Z,0.975} \right],$$

where $q_{Z,0.975}$ is the 97.5'th percentile of $Z$ and

$$\hat{D}_n = \frac{1}{\hat{f}_n(F_n^{-1}(\theta_0))} \left( \frac{4 \theta_0 (1 - \theta_0) \hat{f}_n(F_n^{-1}(\theta_0))}{\hat{g}_n(F_n^{-1}(\theta_0))} \right).$$

**Likelihood ratio estimation of quantiles of $F$**: The next theorem deals with the limiting behavior of the likelihood ratio statistic for testing pointwise hypotheses about quantiles of the survival time distribution.
Theorem 2 Let \(0 < \theta_0 < 1\) be fixed. Consider testing the null hypothesis \(H_0 : F^{-1}(\theta_0) = t_0\) based on current status data. Denote the likelihood ratio statistic for testing \(H_0\) by \(2 \log(\hat{\lambda}_n)\). Suppose that both \(F_0\) and \(G\) (the true distributions of the event time and the observation time) are continuously differentiable with positive derivatives \(f_0\) and \(g\) at \(t_0\). Then under the null hypothesis, 

\[
2 \log(\hat{\lambda}_n) \to_d \int \left\{ (g_{1,1}(z))^2 - (g^0_{1,1}(z))^2 \right\} dz \equiv D. \tag{2}
\]

For a proof of this theorem, see the Appendix.

A confidence region for \(t_0\) can now be obtained by inverting the acceptance region of the likelihood ratio test. This is done in the following way. The LRS is computed under a family of null hypotheses \(F^{-1}(\theta_0) = t\) as \(t\) varies over the line; the 95% confidence region for \(F^{-1}(\theta_0)\) is the set of all values of \(t\) for which the null hypothesis is not rejected. As argued in the proof, the LRS for testing \(F^{-1}(\theta_0) = t\) is identical to the LRS for testing \(F(t) = \theta_0\). Furthermore, it is not difficult to see that the LRS as a function of \(t\) is piecewise constant; for each \(i\), values of \(t \in [T(i), T(i+1))\) yield the same value of the LRS. Hence the exact 95% confidence region for \(F^{-1}(\theta_0)\) can be computed exactly with only finitely many computations (of order at most \(n\); this can be reduced through an intelligent search).

3 Comparisons through simulation studies

In this section we compare and contrast the different methods for the construction of confidence sets described above. We first discuss and illustrate the construction of confidence intervals for \(F\) at a fixed point of interest, and then, more briefly, describe the construction of confidence intervals for the quantiles of \(F\).

3.1 Confidence sets for \(F(t_0)\):

MLE based methods: The major drawback of the MLE based intervals for estimation of \(F\) or \(F^{-1}\) is the need to estimate the densities \(f\) and \(g\). We first focus on nonparametric estimation of \(g\).

Since an i.i.d. sample from \(G\) is available, this can be used to construct the empirical distribution function \(\bar{G}_n\). This has \(n\) support points and kernel smoothing techniques can be applied directly. For the computations in this paper, we estimate \(g\) using a standard normal kernel and bandwidth determined using both likelihood and least squares based cross validation techniques. Least squares based cross–validation is based on minimizing the integrated squared
error loss function $\int (\hat{g}_h(x) - g(x))^2 \, dx$ as a function of $h$ where $\hat{g}_h(x)$ is the kernel density estimator of $g(x)$ using bandwidth $h$. The squared error loss is estimated using leave one out cross validation and results in the following criterion:

$$LSCV(h) = \int \hat{g}_h(x)^2 \, dx - \frac{2}{n-1} \sum_{i=1}^{n} \left( \hat{g}_h(T_i) - \frac{K(0)}{nh} \right)$$

where $K$ is the kernel (see Loader (1999)). For the standard normal kernel the integral on the right side of the above display has a closed form expression in terms of $h$ and $T_1, T_2, \ldots, T_n$.

The optimal $h$ for a given sample $T_1, T_2, \ldots, T_n$ is obtained by minimizing $LSCV(h)$ over a fine grid.

The likelihood based cross-validation technique used in the paper is essentially a variant of the original cross validation technique proposed by Habbema et al. (1974) and Duin (1976). The sample $T_1, T_2, \ldots, T_n$ is divided randomly into roughly two equal parts; call these $D_1$ and $D_2$. Let $\hat{g}_{h,D_i}$ denote the estimate of $g$ using bandwidth $h$ and data $D_i$ (for $i = 1, 2$). Define

$$LCV(h) = \prod_{T_i \in D_1} \hat{g}_{h,D_2}(T_i) \times \prod_{T_i \in D_2} \hat{g}_{h,D_1}(T_i).$$

The optimal bandwidth is chosen by maximizing the above criterion as a function of $h$.

The trickier problem is to estimate the density $f$ at $t_0$; since we never observe the actual failure times (the $X_i$’s) the empirical distribution function for $F$ is not available. A kernel smoothing approach using $F_n$, the MLE of $F$ based on the interval censored data, is the easiest approach. This problem was also raised in Keiding (1991) and the following discussion by Groeneboom (1991). However, as the number of support points for $F_n$ is only $O_p(n^{1/3})$ (resulting in fewer jumps and larger jumps sizes compared to the empirical distribution function that one could construct had the $X_i$’s been known), direct kernel smoothing with naive bandwidth choices may not recover all the information lost in the discrete NPMLE of $F$. Groeneboom (1991) suggested that a bandwidth of order $n^{-1/7}$ (corresponding to the optimal bandwidth for estimating the derivative of a density with a third derivative) would be appropriate in this context, and went on to suggest a bootstrap method with vanishing bootstrap sample size for carrying out the bandwidth selection. Considerations analogous to those of Groeneboom & Jongbloed (2003) might suggest further refinements of these methods. Braun, Duchesne & Stafford (2005) have recently suggested a kernel density method for interval censored data where they extend the usual approach by computing the conditional expectation of the kernel weight corresponding to the $i$’th observation conditional on $I_i$, where $I_i$ is the interval in which the event is known to occur. The estimator is shown to be the solution to a fixed point equation which is then computed using
iterative techniques. For bandwidth selection, cross-validation based ideas are used. However, the computational intensity of this method led us to resort to direct kernel smoothing of the NPMLE. Now, least squares type cross validation does not work in this context, as there is no clear way of writing down the asymptotic mean squared error of a kernel estimator with bandwidth \( h \). On the other hand, likelihood based cross validation can still be employed. For this paper we follow (a version of) the method described in Pan (2000). The data \( \{\Delta_i, T_i\}_{i=1}^n \) are randomly divided into two roughly equal subsets; let \( D_1 \) denote the set of indices corresponding to the first subset and \( D_2 \), the set of indices corresponding to the second. Let \( \hat{f}_{h,D_i}(\cdot) \) denote the density estimator based on the subset based on \( D_i \) using bandwidth \( h \), i.e.

\[
\hat{f}_{h,D_i}(x) = \frac{1}{h} \int K\left(\frac{x - t}{h}\right) d\hat{F}_{D_i}(t),
\]

where \( \hat{F}_{D_i} \) is the NPMLE of \( F \) based on the set \( D_i \). Then the estimate of \( F \) based on \( D_i \) is obtained by integrating \( \hat{f}_{h,D_i} \) and is denoted by \( \hat{F}_{h,D_i} \). Define \( LCV_{cens}(h) \) as:

\[
LCV_{cens}(h) = \prod_{i \in D_1} \hat{F}_{h,D_2}(T_i)^{\Delta_i} \left(1 - \hat{F}_{h,D_2}(T_i)\right)^{1-\Delta_i} \times \prod_{i \in D_2} \hat{F}_{h,D_1}(T_i)^{\Delta_i} \left(1 - \hat{F}_{h,D_1}(T_i)\right)^{1-\Delta_i}.
\]

As for the uncensored case, the optimal bandwidth is chosen by maximizing the above criterion as a function of \( h \). Although likelihood based cross-validation has been criticized for not being efficient enough in a complete data setting (see, for example, Wand & Jones (1995)), it appears that among many of its competitors it is the most reliable in small sampling settings from an extensive simulation study by Grund & Polzehl (1996). Pan (2000) compares the performance of the above kernel estimator to that of the logspline estimator (where the log-density is modelled as a spline) due to Kooperberg & Stone (1992) and concludes that the performances of the two estimators are reasonably comparable. For our simulations, we used a standard normal kernel as before.

**Subsampling based methods:** Subsampling methods provide an alternative way of estimating the quantiles of the limit distribution of the NPMLE. With subsampling based methods, a critical issue is the choice of the block size \( b \). This problem is analogous to the choice of the bandwidth in smoothing problems. Unfortunately, the asymptotic requirements that \( b \to \infty \) and \( b/n \to \infty \) as \( n \to \infty \) give little guidance when faced with a finite sample. Instead, in practice one can use a calibration algorithm as in Delgado, Rodriguez-Poo & Wolf (2001). We present a brief discussion.

Consider the current status model. For a finite sample size \( n \), a subsampling based confidence interval for \( F(t_0) \) (with asymptotic coverage level \( 1 - \alpha \)) will typically not exhibit level exactly
equal to $1 - \alpha$; the actual level will depend upon the underlying distributions $F$ and $G$, the sample size $n$, the point $t_0$ and finally the chosen block size $b$. Indeed, one can think of the actual level $\lambda$ as a function $h$ of $(P_{F,G}, n, t_0, b)$; here $P_{F,G}$ is the distribution of the current status data vector $(\Delta, T)$. The idea now is to adjust the “input” $b$ in order to obtain an actual level close to the nominal one. For fixed $P_{F,G}, n, t_0$, the optimal block size $\hat{b}$ would be the one that minimizes $| h(P_{F,G}, t_0, n, b) - (1 - \alpha) |$; note that $| h(P_{F,G}, t_0, n, b) - (1 - \alpha) | = 0$ may not always have a solution. Analytically $h$ will be extremely complicated to express, but if $F$ and $G$ are known, we can simulate $h(\cdot)$ by generating $n$ i.i.d. observations from the current status model and constructing subsampling based confidence sets for $F(t_0)$ for a number of different block sizes $b$. This process is repeated many times (say $K$) and $h(P_{F,G}, t_0, n, b)$ is estimated as the proportion of subsampling based confidence intervals using block size $b$ that contain $F(t_0)$. One then selects $\hat{b}$ as that $b$ for which $\hat{h}(F, G, t_0, n, b)$ is closest to $1 - \alpha$. However, in reality, $F$ and $G$ are unknown, so the above recipe does not work. However we can replace $P_{F,G}$ by a consistent estimator $\hat{P}_n$; a sensible choice that is always available is the empirical distribution function of the data $\{\Delta_i, T_i\}_{i=1}^n$. We describe the exact algorithm below:

Algorithm for choosing block size:

(a) Fix a selection of reasonable block sizes between limits $b_{low}$ and $b_{up}$.

(b) Generate $K$ pseudo sequences $(\Delta_{k}^*, T_{k}^*)_{k=1}^K$ which are i.i.d. $\hat{P}_n$; with $\hat{P}_n$ equal to the empirical distribution function this amounts to drawing $K$ bootstrap samples from the actual data set.

(c) For each pseudo data set, construct a subsampling based confidence interval (with asymptotic coverage $1 - \alpha$) for $\hat{\theta}_n \equiv F_n(t_0)$ for each block size $b$. Let $I_{k,b}$ be equal to 1, if $\hat{\theta}_n$ lies in the $k$’th interval based on block size $b$ and 0 otherwise.

(d) Compute $\hat{h}(b) = K^{-1} \sum_{i=1}^{K} I_{k,b}$

(e) Find $\hat{b}$ that minimizes $| \hat{h}(b) - (1 - \alpha) |$ and use this as the block size to compute subsampling based confidence intervals based on the original data.

Indeed, this is the algorithm that we use subsequently in analyzing the Rubella data set in
Section 2.3 and the simulation studies to be presented below.

**Likelihood Ratio based method:** The likelihood ratio based method, like the subsampling approach, enables us to find confidence intervals for $F(t_0)$ without having to resort to nuisance parameter estimation. This is because, under the null hypothesis, the likelihood ratio statistic has a universal limit distribution, that is free of the underlying parameters in the problem. Thus one needs to know only the quantiles of the limit distribution $D$; estimates of these quantiles are tabulated in Banerjee & Wellner (2001). To find a confidence set for $\theta = F(t_0)$ one lets $\theta$ vary on a fine grid between 0 and 1 and for each value of $\theta$ computes the likelihood ratio statistic corresponding to the null hypothesis $F(t_0) = \theta$. The $\theta$'s in the confidence set are precisely those for which the null hypothesis fails to be rejected. It may apparently seem that the MLE based method one requires less computation than the likelihood ratio method, since with the former one needs to compute the unconstrained MLE and estimate the constant $C$, using $\hat{C}_n$. However, estimating $\hat{C}_n$ requires estimating $f$ and $g$ and unless decent parametric estimates are available, one needs to resort to maximization of a cross-validation based criterion over a grid and this is computationally very intensive. The major advantage of the the likelihood ratio based method over the MLE based method lies in superior reliability, since it avoids ad-hoc estimation of nuisance parameters completely. Furthermore, the computation with the likelihood ratio method is hardly overwhelming. This is where it enjoys a major advantage over the subsampling based method; the subsampling procedure also becomes computationally very intense as the optimal block size needs to be determined using the bootstrap based method discussed above. This is analogous to the problem of determining the optimal bandwidth with the kernel based method.

**Simulations:** We now present simulation results from the current status model.

**Simulation setting 1:** We took $F = \exp(1), G = \exp(1)$ and $t_0 = \log(2.0)$. Thus $\theta_0 = F(t_0) = 0.5$. We chose sample sizes $n = 50, 75, 100, 200, 500, 800, 1000$. For each value of $n$, we generated 1000 data sets from the current status model and for each data set, we computed 95% confidence intervals for $\theta$, using (i) the MLE based method with nonparametric estimation of $f(t_0)$ and $g(t_0)$, (ii) the likelihood ratio based method, (iii) subsampling based techniques, (iv) parametric (Weibull) based estimation of $f(t_0)$ and $g(t_0)$.

Thus estimates of $f(t_0)$ and $g(t_0)$ were obtained in (i) via kernel smoothing with optimal bandwidth determined through likelihood based cross-validation for both $f$ and $g$ (described above), and in (iv) by fitting Weibull distributions to $F$ and $G$ using maximum likelihood
estimation. and using the resulting parameter MLE’s to estimate \( f(t_0) \) and \( g(t_0) \). For the subsampling based confidence intervals, the optimal block size \( \tilde{b} \) was determined in accordance with the bootstrap based algorithm for selecting block–size, from the following selected block sizes: \( n^{1/3}, n^{1/2}, n^{2/3}, n^{3/4}, n^{0.8}, n^{0.9} \). The average length of the three types of confidence sets was then computed over the 1000 data sets; the proportion of intervals that contained the true parameter value was also recorded for each method.

Table 1 contains the above information. The average length of the confidence sets decreases with increasing sample size \( (n) \) at rate \( n^{-1/3} \). Note that the average length of 95% confidence intervals obtained via the likelihood ratio method (shown in column 2) is smaller than the average length of those computed using the MLE based method with \( f \) and \( g \) estimated as in (i) above, or the subsampling based method (Columns 3 and 4). However, the shortest confidence intervals are obtained using the MLE based method with \( f \) and \( g \) estimated parametrically as in (iv) above (Column 5). A look at the estimated coverage probabilities of these four kinds of intervals (Columns 6, 7, 8 and 9) shows that the MLE based method (i) is anti–conservative for smaller sample sizes, but as the sample size grows the coverage probability improves substantially. The MLE based method (iv) stays substantially anticonservative even at higher sample sizes. The likelihood ratio based C.I.’s are very slightly anti–conservative while the subsampling based C.I.’s are slightly conservative.

MLE based confidence intervals were also computed using least squares based cross–validation for \( g \), to assess the optimal bandwidth. The results (not exhibited here) are very compatible to what is reported in the table above; hence we do not present a separate table. As in the above case, the actual coverage of the 95% MLE based C.I.’s is below or around 90% for smaller sample sizes but improves similar to Column 6 of Table 1, as the sample size increases. These C.I.’s also continue to be slightly wider than their LR based counterparts.

The likelihood ratio and subsampling based techniques therefore seem more reliable for moderate sample sizes than the MLE based ones as far as coverage is concerned. From the point of view of precision it appears that the likelihood ratio based intervals are shorter in average length than the ones based on subsampling without compromising coverage substantially.
Simulation setting 2: The survival time distribution $F$ was changed from Exponential(1) to Gamma(3,1) and the observation time distribution was changed to Uniform(0,5); $t_0$ was taken to be the median value of Gamma(3,1), so $F(t_0) = 0.5$. Likelihood ratio and MLE based confidence intervals were constructed for the same selection of sample sizes as in the previous setting with 1000 data sets being generated for every sample size. For this example, $f(t_0)$ was estimated as before, but kernel smoothing was no longer used to estimate $g(t_0)$. Instead, we used our background knowledge that the observation time distribution is uniform and estimated the density as $(T(n) - T(1))^{-1}$. This gives us a very accurate estimate of $g(t_0)$ even at moderate sample sizes. Table 2 compares the likelihood ratio based confidence intervals to the MLE based ones.

Table 2 around here

Note that the above pattern resembles Simulation 1. The likelihood ratio based intervals outperform the MLE based intervals in terms of both length and coverage, especially at smaller sample sizes where the MLE based intervals are fairly anti-conservative. Subsampling based confidence intervals were not computed for this setting.

Simulation setting 3: The goal of this setting is to understand how the different methods for estimating $F(t_0)$ compare when $t_0$ lies in a region of steep ascent of the distribution function $F$. This is motivated by our analysis of the Rubella data set presented in the next section. The NPMLE of the distribution of time to infection by Rubella jumps from .571 at $t = 12$ years to .857 at $t = 13$ years and stays flat in the interval 13–15 years, thereby indicating that the range 12–13 years is a region of steep ascent of the distribution of time to immunization. The (nonparametric) MLE based C.I. at $t = 12$ years and the subsampling based C.I. at the same time point are seen to be substantially wider than the corresponding C.I.’s computed via other methods (see Figure 2). Also, the (nonparametric) MLE based C.I. for $t = 12$ is the widest C.I. of its type; a similar phenomenon holds the subsampling based C.I.’s. We therefore seek to investigate the relative behavior of the different confidence sets in a region of sudden change of the distribution function.

We chose the survival distribution $F$ as follows: $F(x) = x$ for $x \leq 0.25$, $F(x) = .25 + a(x - .25)^2$ for $.25 < x \leq .25 + \epsilon$, and $F(x) = .75 + \lambda(x - .25 - \epsilon)$ for $.25 + \epsilon < x \leq 1$. Here $a = 20000$, $\epsilon = 1/200$, $\lambda = .25/(.75 - 1/200)$, so that $F(1) = 1$. The distribution function thus constructed is continuous and continuously differentiable in a neighborhood of $t_0 = .25 + \epsilon/2$, where $F(t_0) = .375$. The function $F$ rises very steeply from .25 to .75 in the interval [.25, .25 + $\epsilon$] (so $t_0$ is the midpoint
of this interval) and the slope of \( F \) at \( t_0 \) is equal to 100. The observation time distribution was taken to be uniform on the interval \((0, 1)\). Sample sizes \( n = 50, 100, 200 \) were considered and 1000 data sets were generated for each \( n \). For the MLE based method \( f \) was estimated using kernel smoothing in the usual manner and \( g(t_0) \) was estimated as \( 1/(T(n) - T(1)) \), as in Simulation 2. For the subsampling based method, the same selection of block sizes was used as in Simulations 1 and 2; \( K = 300 \) bootstrap samples were drawn for determining the optimal block size, and the final (symmetric) subsampling based intervals were constructed using \( S = 300 \) subsamples of the optimal block size. We present the results below in Table 3.

Table 3 shows how the different methods adapt to a sudden change in \( F \). The likelihood ratio method continues to be stable, producing confidence sets with coverage close to the nominal (and at modest sample sizes). The MLE and the subsampling based methods however do not react well at all. While they produce shorter confidence sets on an average (than the likelihood ratio based method), this is only at the expense of extremely erroneous coverage. Hence, in a situation of this sort, the likelihood ratio method seems to be the only reliable candidate.

### 3.2 Confidence sets for \( F^{-1}(\theta_0) \)

We illustrate the construction of confidence sets for the \( \theta_0 \)'th quantile of \( F \), for a given \( \theta_0 \) strictly between 0 and 1. Confidence intervals for the quantile can be constructed, using either the likelihood ratio based approach (Theorem 2) or the MLE based approach (Theorem 1). The latter requires estimation of \( f(t_0) \) and \( f(t_0) \) where \( t_0 \equiv F^{-1}(\theta_0) \); in the following simulation setting, this is done using kernel density estimates \( \hat{f} \) and \( \hat{g} \) with optimal bandwidth \( h \) determined by likelihood based cross-validation for \( f \) and least squares based cross-validation for \( g \); the point \( t_0 \) is estimated by \( F_n^{-1}(\theta_0) \).

Table 4 shows how the different methods adapt to a sudden change in \( F \). The likelihood ratio method continues to be stable, producing confidence sets with coverage close to the nominal (and at modest sample sizes). The MLE and the subsampling based methods however do not react well at all. While they produce shorter confidence sets on an average (than the likelihood ratio based method), this is only at the expense of extremely erroneous coverage. Hence, in a situation of this sort, the likelihood ratio method seems to be the only reliable candidate.

### Simulations:

We present results from a simulation study involving estimation of the median of the Exponential(1) distribution, based on a current status model. We compare the likelihood ratio based method to the MLE based method. The true value of the median is \( \log 2 = 0.693 \).

The observation time \( G \) is also chosen to be Exponential (1), as in Simulation 1. The sample size is allowed to vary as in the previous simulation settings; for each \( n \), 1000 data sets of size...
are generated and asymptotically 95% confidence sets for $F^{-1}(0.5)$ are obtained using the
two methods. For the MLE based method $f$ and $g$ are estimated using kernel smoothing, with
likelihood based CV being employed to select the optimal bandwidth for $f$ and least squares based
CV being used to select the optimal bandwidth for $g$. Table 4 shows the average lengths of the
confidence intervals obtained using these methods (over 1000 replicates) and the corresponding
coverage probabilities. The observed coverage probabilities for the likelihood ratio based method
are seen to be close to 95%; however, the MLE based confidence intervals are once again fairly
anti-conservative, and continue to stay so even at higher sample sizes (unlike what happens when
estimating the distribution function itself). The likelihood ratio based intervals are seen to be
somewhat wider than the MLE based ones, but are more reliable, since the reduced widths of the
MLE based intervals come at the expense of coverage.

4 Analyzing the Austrian Rubella Data

The methods explained in the above section are illustrated here on some data made available to us
by Neils Keiding. The data set concerns 230 Austrian males older than 3 months for whom the exact
date of birth was known. Each individual was tested at the Institute of Virology, Vienna during
the period 1-25 March 1988 for immunization against rubella. The Austrian vaccination policy
against rubella had then for some time been to routinely immunize girls just before puberty but
not to vaccinate the males, so that the males can be taken to represent an unvaccinated population.

The goal here is to estimate the distribution of the time to infection (and subsequent immunization)
by rubella in the male population. It is assumed that immunization, once achieved, is lifelong.
We denote the distribution of the time to immunization by $F$. Let $T_i$ denote the age of the $i$'th
individual at the time of testing for immunization and let $X_i$ denote the time to immunization.
Thus $X_i$ is distributed as $F$. Let $G$ denote the age distribution in the population. Thus $T_i$ is
distributed as $G$. However $X_i$ is not observed; the only information available is whether the person
is immunized or not at the time of testing. Thus we observe $(\Delta_i = 1\{X_i \leq T_i\},T_i)$. Under the
i.i.d. assumption on $(X_i,T_i)_{i=1}^n$ and the assumption of independence between $X_i$ and $T_i$ (current
age and age of immunization are independent) we are in the current status framework.

Keiding, Begtrup, Scheike & Hasibeder (1996) analyzed this data using the current status
model. They used the asymptotic distribution of the MLE to obtain (pointwise) confidence sets
for $F$. The densities at the point of interest $t_0$, namely $f(t_0)$ and $g(t_0)$, were estimated using
parametric methods. In particular \( g(t_0) \) was estimated by fitting a Weibull distribution to \( G \). The parameters of the Weibull were obtained by using maximum likelihood with the observed \( T_i \)'s as data. On the other hand \( f(t_0) \) was estimated as,

\[
\hat{f}(t_0) = (1 - \hat{F}_n(t_0)) \hat{\lambda}_f(t_0),
\]

where \( \hat{F}_n \) is the NPMLE of \( F \) and \( \hat{\lambda}_f \) is an estimate of the instantaneous hazard function \( \lambda_f \) corresponding to \( F \); in terms of the Weibull parameters \( \alpha \) and \( \beta \), \( \lambda_f \) is given by: 

\[
\lambda_f(x) = \alpha \beta x^{\beta-1}.
\]

For details on the estimation of \( \lambda_f \), we refer the reader to pages 121-123 of Keiding et al. (1996).

Figure 1 around here

Figure 1 shows the NPMLE of the distribution of the time to immunization along with likelihood ratio based 95% confidence intervals for \( F(t) \) with \( t \) varying over the sequence 1, 2, \ldots, 75 years. The distribution function is seen to rise steeply in the age–range 0–20 years, with a dramatic jump from 12 to 13 years of age. There is no significant change beyond 30 years, indicating that almost all individuals are immunized in their youth.

To compare the several types of confidence intervals for this data we computed: (a) the likelihood ratio based C.I.'s for different ages; (b) MLE based C.I.'s constructed using the parametric Weibull fits for \( F \) and \( G \) to estimate \( C_n \); (c) MLE based C.I.'s constructed using kernel smoothing procedures to estimate \( C_n \) (the optimal bandwidth for \( f \) is chosen using likelihood based CV and that for \( g \) is determined through least squares based CV); and (d) subsampling based confidence intervals which are constructed in the exact same way as in Simulations 1 and 3.

Figure 2 around here

Figure 2 gives a comparison of the lengths of the confidence intervals computed via the four different methods. The likelihood ratio based confidence intervals are seen to be generally shorter than the subsampling based intervals (with a few exceptions in the age range 26–30). The likelihood ratio based intervals are regularly behaved in the sense that their left end points and right end points are monotone increasing in \( t \). While this is not a requirement (since we are not constructing confidence bands) this certainly seems to be desirable, at least from an aesthetic perspective. This property is not shared by any of the MLE based C.I.'s; in particular, the confidence intervals obtained via kernel smoothing exhibit the most pronounced violations. In fact, they react very sharply to rapid changes in the distribution function. The NPMLE of \( F \) jumps
from .571 at \( t = 12 \) to .857 at \( t = 13 \); this is the most dramatic change in \( F \) over its entire domain and suggests that the age range 12 – 13 is a zone of steep ascent of \( F \). This is reflected very heavily in the corresponding confidence intervals via kernel smoothing for \( t = 12, 13 \). The confidence intervals are much wider at these points (especially the C.I. for \( t = 12 \), which is the widest C.I. via kernel smoothing) as compared to the C.I.’s for neighboring time points. This is also the case but to a lesser extent for the subsampling based C.I.’s (the C.I. for \( t = 12 \) is the widest C.I.). On the other hand, the MLE based C.I.’s obtained using parametric fits and the likelihood ratio based C.I.’s do not react so drastically to the jump. Our simulation results indicate that in a zone of rapid increase of the distribution function the likelihood ratio method is substantially more reliable than other methods (unless good parametric fits to the data are available).

It is also clear that none of the methods can be expected to come up with the shortest intervals in any given situation. The C.I.’s via kernel smoothing tend to be shorter than the likelihood ratio based C.I.’s on an overall basis, but this is not necessarily a virtue, as our simulation studies indicate that the reduced length of the MLE based intervals may often be associated with suboptimal coverage. From our experience, it seems that the likelihood ratio based intervals adapt nicely across different situations in the sense that the length is adjusted optimally to maintain coverage close to the nominal. This adaptability is not exhibited by the MLE based confidence intervals.

From Figure 2 we see that the lengths of the different C.I.’s more or less agree at the very beginning (with the subsampling intervals being shortest), then in the range from 2-8 years or so, both the LR and the the two other nonparametric intervals become shorter, but the nonparametric-MLE intervals become too short. On the other hand while the “parametrically based” intervals decrease slowly in length, and do not react to the steep change in the distribution function in the interval from 8-14 years, all three of the nonparametric methods try to “catch” the rapid increase of \( F \) in the interval from 8-14 years. The nonparametric-MLE method “overshoots” dramatically while the LR and subsampling approaches seem to be closer to each other, with the LR intervals being shorter. After 14 years or so the nonparametric-MLE intervals are definitely shorter (with the exception of 23 years). It looks as if the parametric - mle method gives shorter intervals than the LR and sub-sampling methods from 8 years on, but probably does not have good coverage properties in view of the result from Simulation 1 given in Table 1. Of course this is the difficulty with comparisons in this real example: the actual coverage probabilities are unknown.

Quantile estimation and the rubella data set: We apply the two different methods of
quantile estimation discussed above to the rubella data set. We estimate selected quantiles of the
distribution of the time to immunization.

Table 5 around here

Table 5 exhibits confidence intervals for \( F^{-1}(p) \) for \( p \) varying across the sequence .40, .60, .70, .90,
obtained by three different methods. Column 1 displays the different values of \( p \), Column 2 the
values of \( F_n^{-1}(p) \) and Column 3 shows the MLE based confidence intervals estimated using the
Weibull parametric fits for \( f \) and \( g \) as obtained in Keiding, Begtrup, Scheike, and Hasibeder
(1996). Recall that by Theorem 1, an approximate 95% confidence interval for \( F^{-1}(p) \) is given by

\[
\left[ F_n^{-1}(p) - \hat{D}_n n^{-1/3} q_{Z,.975}, \ F_n^{-1}(p) + \hat{D}_n n^{-1/3} q_{Z,.975} \right],
\]

where \( q_{Z,.975} \) is the 97.5'th percentile of \( Z \) and \( \hat{D}_n \) estimates

\[
D \equiv \frac{1}{f(F^{-1}(p))} \left( \frac{4 f(F^{-1}(p)) p (1 - p)}{g(F^{-1}(p))} \right). 
\]

We estimate \( g(F^{-1}(p)) \) by \( \hat{g}(F_n^{-1}(p)) \), where \( \hat{g}(\cdot) \) is the Weibull based estimate of \( g(\cdot) \) (with
Weibull parameters obtained from Keiding et al. (1996). To estimate \( f(F^{-1}(p)) \) we note that
this is equal to \( (1 - p) \lambda_f(F^{-1}(p)) \), where \( \lambda_f \) is the instantaneous hazard corresponding to \( f \).
We estimate \( \lambda_f(F^{-1}(p)) \) by \( \hat{\lambda}_f(F_n^{-1}(p)) \), where \( \hat{\lambda}_f \) is the Weibull based estimate of \( \lambda_f \) (with
the Weibull parameters obtained from Keiding et al. (1996)). Column 4 shows the MLE based confidence intervals where \( D \) is estimated nonparametrically using kernel based estimates of \( f \) and \( g \); the point \( F^{-1}(p) \) is estimated as before by \( F_n^{-1}(p) \). The optimal bandwidth for estimating \( f \) is
chosen using likelihood based CV, while that for estimating \( g \) is chosen using least squares based
CV. Finally Column 5 shows the likelihood ratio based C.I.'s for \( F^{-1}(p) \).

One pleasing property of the LR based confidence intervals in Figure 1 is the monotonicity
of the left end-points as well as the right end-points with increasing \( t \) and in Table 5, with
increasing \( p \). This is not exhibited by the MLE based confidence intervals. Since we are only
concerned with pointwise confidence sets here, monotonicity is not a crucial statistical issue but if
the question was one of constructing confidence bands, it would be a key requirement.

ACKNOWLEDGEMENTS: Thanks to Piet Groeneboom for many discussions about
interval censoring, to Niels Keiding for providing the rubella data, to Michael Wolf for some
extremely useful pointers and references about sub-sampling, and to Michael Woodroofe for
suggesting the quantile estimation problem. M. Banerjee’s research for this paper was supported by an NSF Grant; J. A. Wellner’s research was supported grants from the NSF and NIAID.

References


---

**Moulinath Banerjee**
**DEPARTMENT OF STATISTICS**
439, WEST HALL
550, E. UNIVERSITY
ANN ARBOR, MICHIGAN 48109

*e-mail: moulib@umich.edu*

5 Appendix

**Proof of Lemma 1**: Clearly

\[ A_{n,\lambda,x} \equiv \left\{ F_n^{-1}(\theta_0 + \lambda n^{-1/3}) \leq F^{-1}(\theta_0) + x n^{-1/3} \right\}, \]
which equals the event
\[ \left\{ F_n \left( F^{-1}(\theta_0) + x n^{-1/3} \right) \geq \theta_0 + \lambda n^{-1/3} \right\} \]

using the equivalence that for any distribution function $H$ and $u \in (0, 1)$ and $t$ real, $H^{-1}(u) \leq t$ if and only if $H(t) \geq u$. Now, $F^{-1}(\theta_0) = t_0$. Also, since $F$ has a continuous positive derivative in a neighborhood of $t_0$, it follows that $\theta_0 = F(t_0)$. Then, the above display can be rewritten as
\[ \left\{ n^{1/3} \left( F_n(t_0 + x n^{-1/3}) - F(t_0) \right) \geq \lambda \right\}. \]

From Theorem 2.4, page 1710 of Banerjee and Wellner (2001) we know that
\[ n^{1/3} \left( F_n(t_0 + x n^{-1/3}) - F(t_0) \right) \to_d \frac{1}{g(t_0)} g_{a,b}(x), \]

where $a = \sqrt{g(t_0) \theta_0 (1 - \theta_0)}$ and $b = f(t_0) g(t_0)/2$ and $g_{a,b}$ is the right-derivative of the greatest convex minorant of the process $X_{a,b}(t)$ (defined in Section 1 at the point $x$). Therefore
\[ \lim_{n \to \infty} P \left( n^{1/3} \left( F_n(t_0 + x n^{-1/3}) - F(t_0) \right) \geq \lambda \right) = P \left( \frac{1}{g(t_0)} g_{a,b}(x) \geq \lambda \right). \quad (3) \]

Now, using the switching relationship (see e.g. Banerjee (2000), Lemma 3.6.11, page 144) it follows that
\[ P \left( \frac{1}{g(t_0)} g_{a,b}(x) \geq \lambda \right) = P \left( \text{argmin}_h X_{a,b}(h) - g(t_0) \lambda h \leq x \right). \]

But
\[ X_{a,b}(h) - g(t_0) \lambda h = P \left( \frac{a}{b} \right)^{2/3} \text{argmin}_h X_{1,1}(h) + (1/2) \left( \lambda g(t_0)/b \right), \]

by e.g. van der Vaart and Wellner (1996), problem 5, page 308. Hence, the limiting probability in (3) is
\[ P \left( \left( \frac{a}{b} \right)^{2/3} \text{argmin}_h X_{1,1}(h) + \lambda/f(t_0) \leq x \right). \]

But $\text{argmin}_h X_{1,1}(h) \equiv Z$. We thus have
\[ \lim_{n \to \infty} P \left( n^{1/3} \left( F_n^{-1}(\theta_0 + \lambda n^{-1/3}) - F^{-1}(\theta_0) \right) \leq x \right) = P \left( \left( \frac{a}{b} \right)^{2/3} Z + \lambda/f(t_0) \leq x \right). \]

Noting that
\[ \left( \frac{a}{b} \right)^{2/3} = \left( \frac{4 \theta_0 (1 - \theta_0)}{g(t_0) f(t_0)^2} \right)^{1/3} \]

finishes the proof. \( \Box \)

**Proof of Theorem 2:** Since the underlying distribution function is continuous, with probability 1 it is the case that
\[ 0 < T_1 < T_2 < \ldots T_m < t_0 < T_{m+1} < \ldots < T_n, \]

23
where $T(i)$’s are the ordered observation times and $m$ is the number of observation times not exceeding $t_0$. We shall show that

$$\sup_{F^{-1}(\theta_0) = t_0} L_n(F) = \sup_{F(t_0) = \theta_0} L_n(F), \tag{4}$$

where $L_n(F)$ as before, denotes the log-likelihood for $n$ observations in the interval censoring model. Now,

$$\tilde{\lambda}_n \equiv \frac{\sup_{F} L_n(F)}{\sup_{F^{-1}(\theta_0) = t_0} L_n(F)} ,$$

and let

$$\lambda_n = \frac{\sup_{F} L_n(F)}{\sup_{F(t_0) = \theta_0} L_n(F)} .$$

It then follows immediately that

$$2 \log \tilde{\lambda}_n = 2 \log \lambda_n .$$

But $2 \log \lambda_n$ is the likelihood ratio statistic computed under the null hypothesis $F(t_0) = \theta_0$. Since $F_0$ is continuous with a positive density at $t_0$ and $F_0^{-1}(\theta_0) = t_0$ under the null, clearly it is the case that $F_0(t_0) = \theta_0$. A direct appeal to Theorem 2.5 of Banerjee and Wellner (2001) then shows that

$$2 \log \tilde{\lambda}_n \rightarrow_d \int \left( (g_{1,1}(z))^2 - (g_{0,1}(z))^2 \right) \, dz .$$

To prove (4) we proceed as follows. We note that

$$F^{-1}(\theta_0) = t_0$$

if and only if,

$$F(x) < \theta_0 \leq F(t_0) \ \forall \ x < t_0 . \tag{5}$$

Recalling that

$$L_n(F) = \sum_{i=1}^{n} \left( \delta(i) \log F(T(i)) + (1 - \delta(i)) \log (1 - F(T(i))) \right) ,$$

we see that the supremum of $L_n(F)$ under $F^{-1}(\theta_0) = t_0$ is precisely the supremum of

$$\phi(w) = \sum_{i=1}^{n} \left( \delta(i) \log w_i + (1 - \delta(i)) \log (1 - w_i) \right) ,$$

over the set $0 \leq w_1 \leq w_2 \ldots \leq w_m < \theta_0 \leq w_{m+1} \leq \ldots \leq w_n \leq 1$ (since $T(m) < t_0$ with probability 1, under the null hypothesis that $F^{-1}(t_0) = \theta_0$, it must be the case (by (5) that $w_m \equiv F(T(m)) < \theta_0$). Clearly, the supremum of $\phi(w)$ over this set cannot be larger than the supremum of $\phi(w)$ over the set

$$0 \leq w_1 \leq w_2 \ldots \leq w_m \leq \theta_0 \leq w_{m+1} \leq \ldots \leq w_n \leq 1 .$$
But we know that the supremum of \( \phi(w) \) over this set is indeed attained for a \( \tilde{w} \) and is precisely 

\[
\sup_{F(t_0) = \theta_0} L_n(F).
\]

If \( \tilde{w}_m \) is strictly less than \( \theta_0 \), then of course \( \tilde{w} \) is the maximizer of \( \phi(w) \) over the set \( 0 \leq w_1 \leq w_2 \ldots \leq w_m < \theta_0 \leq w_{(m+1)} \leq \ldots \leq w_n \leq 1 \) as well and (4) is trivially satisfied. In case \( \tilde{w}_m \) is equal to \( \theta_0 \), there exists a smallest \( k \leq m \) such that \( \tilde{w}_{(k-1)} < \theta_0 \) and \( \tilde{w}_{(k)} = \ldots = \tilde{w}_{(m)} = \theta_0 \) in which case, we define for each \( N \), the vector \( \tilde{w}^N \) by setting

\[
\tilde{w}_{(k)}^N = \ldots = \tilde{w}_{(m)}^N = \theta_0 - \frac{1}{N}
\]

and letting \( \tilde{w}_{(i)}^N = \tilde{w}_{(i)} \) for all other \( i \)'s. Clearly, for all sufficiently large \( N \), each \( \tilde{w}^N \) is in the set \( 0 \leq w_1 \leq w_2 \ldots \leq w_m < \theta_0 \leq w_{(m+1)} \leq \ldots \leq w_n \leq 1 \) and using the continuity of \( \phi \),

\[
\lim_{N \to \infty} \phi(\tilde{w}^N) \to \phi(\tilde{w}).
\]

This immediately implies that

\[
\sup_{F(t_0) = \theta_0} L_n(F) = \phi(\tilde{w}) = \sup_{F^{-1}(\theta_0) = t_0} L_n(F).
\]

This finishes the proof. \( \square \)

**Comment:** The above proof shows that there may not exist any distribution function \( F \), in the null hypothesis \( H_0 \) specified by \( F^{-1}(\theta_0) = t_0 \), for which the supremum of \( L_n(F) \) over all distributions in the null hypothesis is attained. If the MLE of \( F \) under the constraint \( F(t_0) = \theta_0 \), which we denote by \( F^0_n \) as before, satisfies \( F^0_n(T_{(m)}) < \theta_0 \), then indeed, the supremum of \( L_n(F) \) over all \( F \) in \( H_0 \) is attained at \( F^0_n \); if however \( F^0_n(T_{(m)}) = \theta_0 \), then there exists no maximizer; however we can approach the supremum through a sequence of distribution functions \( G_N \) in \( H_0 \) where \( G_N(T_{(i)}) = \tilde{w}^N_i \).
Table 1: Comparing confidence intervals, Simulation 1.

(“len”=Observed average Length; “cv”=Observed average Coverage)

<table>
<thead>
<tr>
<th>n</th>
<th>len(lrt)</th>
<th>len(mle)</th>
<th>len(sub)</th>
<th>len(par-mle)</th>
<th>cv(lrt)</th>
<th>cv(mle)</th>
<th>cv(sub)</th>
<th>cv(par-mle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.485</td>
<td>0.515</td>
<td>.587</td>
<td>.412</td>
<td>.942</td>
<td>.868</td>
<td>.925</td>
<td>.805</td>
</tr>
<tr>
<td>75</td>
<td>0.428</td>
<td>0.463</td>
<td>.516</td>
<td>.367</td>
<td>.943</td>
<td>.898</td>
<td>.950</td>
<td>.810</td>
</tr>
<tr>
<td>100</td>
<td>0.390</td>
<td>0.420</td>
<td>.465</td>
<td>.336</td>
<td>.940</td>
<td>.889</td>
<td>.959</td>
<td>.839</td>
</tr>
<tr>
<td>200</td>
<td>0.308</td>
<td>0.326</td>
<td>.370</td>
<td>.269</td>
<td>.941</td>
<td>.910</td>
<td>.948</td>
<td>.860</td>
</tr>
<tr>
<td>500</td>
<td>0.231</td>
<td>0.243</td>
<td>.282</td>
<td>.198</td>
<td>.946</td>
<td>.927</td>
<td>.956</td>
<td>.861</td>
</tr>
<tr>
<td>800</td>
<td>0.198</td>
<td>0.210</td>
<td>.235</td>
<td>.170</td>
<td>.949</td>
<td>.936</td>
<td>.953</td>
<td>.851</td>
</tr>
<tr>
<td>1000</td>
<td>0.183</td>
<td>0.190</td>
<td>.226</td>
<td>.158</td>
<td>.941</td>
<td>.910</td>
<td>.951</td>
<td>.874</td>
</tr>
</tbody>
</table>
Table 2: Comparing confidence intervals, Simulation 2.

(“len”=Observed average Length; “cv”=Observed average Coverage)

<table>
<thead>
<tr>
<th>$n$</th>
<th>len(lrt)</th>
<th>len(mle)</th>
<th>cv(lrt)</th>
<th>cv(mle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.501</td>
<td>0.518</td>
<td>.938</td>
<td>.839</td>
</tr>
<tr>
<td>75</td>
<td>0.450</td>
<td>0.468</td>
<td>.946</td>
<td>.888</td>
</tr>
<tr>
<td>100</td>
<td>0.416</td>
<td>0.433</td>
<td>.952</td>
<td>.890</td>
</tr>
<tr>
<td>200</td>
<td>0.333</td>
<td>0.352</td>
<td>.949</td>
<td>.929</td>
</tr>
<tr>
<td>500</td>
<td>0.243</td>
<td>0.262</td>
<td>.954</td>
<td>.942</td>
</tr>
<tr>
<td>800</td>
<td>0.212</td>
<td>0.224</td>
<td>.947</td>
<td>.925</td>
</tr>
<tr>
<td>1000</td>
<td>0.195</td>
<td>0.210</td>
<td>.963</td>
<td>.941</td>
</tr>
</tbody>
</table>
Table 3: Comparing confidence intervals, Simulation 3.

("len"=Observed average Length; "cv"=Observed average Coverage)

<table>
<thead>
<tr>
<th>n</th>
<th>len(lrt)</th>
<th>len(mle)</th>
<th>len(sub)</th>
<th>cv(lrt)</th>
<th>cv(mle)</th>
<th>cv(sub)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.720</td>
<td>0.506</td>
<td>.380</td>
<td>.968</td>
<td>.632</td>
<td>.443</td>
</tr>
<tr>
<td>100</td>
<td>0.658</td>
<td>.513</td>
<td>.363</td>
<td>.965</td>
<td>.706</td>
<td>.510</td>
</tr>
<tr>
<td>200</td>
<td>0.608</td>
<td>0.460</td>
<td>.345</td>
<td>.949</td>
<td>.707</td>
<td>.589</td>
</tr>
</tbody>
</table>
Table 4: C.I.'s for the median using the LR and MLE based methods.

(“len”=Observed average Length; “cv”=Observed average Coverage)

<table>
<thead>
<tr>
<th>n</th>
<th>len (lrt)</th>
<th>len (mle)</th>
<th>cv(lrt)</th>
<th>cv(mle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.962</td>
<td>.933</td>
<td>0.938</td>
<td>.854</td>
</tr>
<tr>
<td>75</td>
<td>0.871</td>
<td>.868</td>
<td>0.935</td>
<td>.830</td>
</tr>
<tr>
<td>100</td>
<td>0.788</td>
<td>.707</td>
<td>0.935</td>
<td>.856</td>
</tr>
<tr>
<td>200</td>
<td>0.626</td>
<td>.556</td>
<td>0.941</td>
<td>.838</td>
</tr>
<tr>
<td>500</td>
<td>0.470</td>
<td>.410</td>
<td>0.957</td>
<td>.834</td>
</tr>
<tr>
<td>800</td>
<td>0.407</td>
<td>.373</td>
<td>0.952</td>
<td>.808</td>
</tr>
<tr>
<td>1000</td>
<td>0.367</td>
<td>.261</td>
<td>0.950</td>
<td>.828</td>
</tr>
</tbody>
</table>
Figure 1: Likelihood ratio based confidence intervals.
Figure 2: Comparison of confidence interval lengths, Rubella data:
diamond = LR; star = MLE(par); square = MLE(nonpar); triangle = MLE(sub)
Table 5: Confidence intervals for the quantiles of the time to immunization by rubella using different methods.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\bar{F}_n^{-1}(p)$</th>
<th>MLE based (p)</th>
<th>MLE based (np)</th>
<th>LR based.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.40</td>
<td>1.2301</td>
<td>0.0-3.9648</td>
<td>0.0-3.2914</td>
<td>0.5288-9.3425</td>
</tr>
<tr>
<td>0.50</td>
<td>5.6411</td>
<td>0.8887 - 10.3935</td>
<td>3.5516-7.7306</td>
<td>0.5342 - 12.4877</td>
</tr>
<tr>
<td>0.90</td>
<td>23.2219</td>
<td>6.1027 - 40.3411</td>
<td>21.7355-24.7083</td>
<td>12.4274 - 34.789</td>
</tr>
</tbody>
</table>