THESIS FOR THE DEGREE OF LICENTIATE OF TECHNOLOGY

On Bootstrapping Survival data

Magnus Åstrand

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY and GÖTEBORG UNIVERSITY
Göteborg, Sweden 2000

On Bootstrapping Survival data
MAGNUS ÅSTRAND

ISSN 0347–2809/NO 2000:64
Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31–772 1000

Chalmers University of Technology
Göteborg, Sweden 2000

Abstract

Many statistical methods are based on the crucial assumption that all observations are stochastically independent. In situations when this assumption does not hold the methods of analysing data often gets very complicated and the options fewer. When estimating the survival function for survival data arising from the random censorship model the Kaplan-Meier estimator is very often used. Under the assumption of observations being stochastically independent it has been showed to enjoy many nice properties. It has also been showed that the Kaplan-Meier estimator still is consistent when groups of observations are dependent. In this thesis we present a method for constructing confidence intervals for the survival function from independent observations based on the Kaplan-Meier estimator. The method resembles likelihood intervals, but instead of relying on an asymptotic distribution of the likelihood ratio, the intervals are based on the exact distribution calculated by means of bootstrap. Different ways of expanding the method to the case of groups of observations being dependent are discussed.

Keywords: Bootstrap, Kaplan-Meier, Survival, Semiparametric, Exact, Dependent.

MSC2000: 62N01, 62N02, 62P10, 62H10, 62H12, 62G05, 62G09.

Acknowledgements

To my supervisor Professor Sture Holm for your ideas and optimism, my former and present roommate Dan and Sara for putting up with my ever so happy mood, my Annika for simply being the lovely person that you are, and to everyone who has helped and inspired me in various ways.

Thank you!

Magnus Åstrand
Göteborg, November 2000

Contents

Chapter 1

Introduction

Consider the random censorship model, i.e. we have n identically and independent failure times distributed according to the survival function F. Each failure time is subject to right censoring by the censor variable, which is assumed stochastically independent of the failure times and distributed according to the survival function G. We do not observe both the failure time and the censor variable for each pair, but rather the minimum and an indicator of which one was the smallest. Our goal is to estimate the survival function F of the failure times. In this situation the Kaplan-Meier estimator (KM), also called the product limit estimator, of F, (Böhmer, 1912), (Kaplan & Meier, 1958), is very often used and has many nice properties.

Akritas (1986) studied different ways of bootstrapping the KM estimator. The two methods are resampling cases, (Efron, 1981a), and resampling from the KM estimator, (Reid, 1981). Akritas studied the asymptotic behaviour of the KM estimator and confidence bands for the survival function F. He showed that Reid's proposal does not give asymptotically correct confidence bands. A third option is conditional bootstrapping, (Hjort, 1985), which was studied by Kim (1995). He concluded that conditional bootstrapping estimates the conditional variance of the KM estimator better than Efron's approach.

Ying and Wei (1994) studied the behaviour of KM when the failure times no longer are mutually independent. They concluded that the KM estimator is still consistent under rather mild assumptions. For highly stratified data these assumptions where trivially correct. Confidence intervals for F using a modified Greenwood formula were considered by Eriksson and Adell (1994). They showed that falsely assuming independence could have serious effect on the confidence level of the interval. However, they had a fix censoring point in their simulations and did not study the KM estimator at points beyond this. This means that all survival estimators are frequencies of overdispersed data.

The aim of this thesis is to study confidence intervals for the survival rate based on the KM estimator. The first section of chapter 2 contains a short introduction to bootstrap and how one may use this in order to find confidence intervals. After the introduction to bootstrap we start with the simplest possible case of survival data, no censored observations and all failure times independent. In this situation the KM estimator equals the relative frequency of survivals, thus the model to consider is the binomial model. Several confidence intervals obtained by sampling cases are investigated, i.e. the "classic" way of bootstrapping. We use the exact binomial confidence interval as the golden standard. In the same chapter we also compare different ways of bootstrapping when groups of observations are dependent. They are compared with respect to the ratio expected variance to the true variance. We still assume there are no censored observations, hence we have an overdispersed binomial model.

In the next chapter we proceed with independent but possibly censored observations. We introduce a semiparametric confidence interval for the survival rate. The interval is computed using bootstrap and in case of no censored observations it coincides with the exact binomial interval. The most promising classic bootstrap interval of chapter 2 is the accelerated bias corrected percentile interval. How this is computed for censored observations is also shown. These two intervals together with a likelihood ratio interval are compared in two simulations.

In chapter 4 we relax the assumption of independence between all failure times, by allowing dependence within groups of observations. In this situation the ordinary KM estimator (ignoring the dependence) is still a valid estimator. An example where data can be considered arising from this kind of model is the lifetimes of several individuals who are related in some way and therefore can not be treated as independent. Another example is the failure time of several similar organs of the same individual. The specific example that was the basis for the work of this thesis comes from oral surgery for the implantation of fixtures to support dental prosthesis. One individual has several fixtures implanted and the time until failure of these fixtures may be dependent. In chapter 4 we discuss different ways of bootstrapping from this kind of data.

Chapter 2

Bootstrapping binomial data

If there are no censored observations the Kaplan-Meier estimator is reduced to the relative frequency of survivals. Thus, when studying different ways of forming confidence intervals for the survival function, a good start is to study binomial data.

In this chapter we will start by considering the binomial model. Then we will continue with an overdispersed binomial model. The latter model corresponds to the kind of dependency structure of the survival model in chapter 4. But first of all, we start with a short introduction to bootstrap and different ways of finding confidence limits using bootstrap.

2.1 Bootstrap in general

Let $x_1, \ldots, x_n$ be a sample of independent random variables all distributed according to the same distribution function F. Let θ be the parameter that we wish to estimate. The parameter θ is a function of F, e.g. the expected value of F. Suppose that we have an estimator $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ of θ at hand. Moreover, besides estimating θ we also want to find a confidence interval.

In order to do this we need to know the distribution function F, or find a pivot statistic with known distribution under some model assumptions, or use some other trick. The simple idea of bootstrap is to solve this problem by replacing the unknown distribution F with the empirical distribution,

$$F_n(x) = \frac{\#\{i; x_i \leq x\}}{n}$$

determined by the sample.

Thus, we can by means of simulation calculate the distribution of the estimator under F_n:

1. Simulate a bootstrap sample of n observations, $x_1^*, \ldots, x_n^*$ [1] from the empirical distribution F_n. Since the empirical distribution function places mass n^{-1} on each of the n observation of the original sample, we may just draw observations from $x_1, \ldots, x_n$ at random with replacement.

2. Using the same function as on the original sample, but now on the simulated sample, calculate the estimator: $\hat{\theta}^* = \hat{\theta}(x_1^*, \ldots, x_n^*)$.

These two steps are repeated N times, and by using the obtained values we get a picture of the distribution of $\hat{\theta}^*$. In particular we can find L and U such that

$$\mathsf{P}^* \left(L \leq \hat{\theta}^* - \hat{\theta} \leq U \right) \approx 1 - 2\alpha.$$

If we let $\hat{\theta}^*_{(\alpha)}$ and $\hat{\theta}^*_{(1-\alpha)}$ denote the αN'th and $(1-\alpha)N$'th ordered values of $\hat{\theta}^*$ respectively, we may use $L = \hat{\theta}^*_{(\alpha)} - \hat{\theta}$ and $U = \hat{\theta}^*_{(1-\alpha)} - \hat{\theta}$.

2.1.1 Confidence intervals

One way of calculating confidence intervals, which originates from Efron (1979), is to approximate the distribution of $\hat{\theta} - \theta$ with the distribution of $\hat{\theta}^* - \hat{\theta}$ observed in the simulation. Thus, we assume that

$$\mathsf{P} \left(L \leq \hat{\theta} - \theta \leq U \right) \approx 1 - 2\alpha \tag{2.1}$$

holds for every θ. Using this we get the *basic* confidence interval:

$$[2\hat{\theta} - \hat{\theta}^*_{(1-\alpha)}, 2\hat{\theta} - \hat{\theta}^*_{(\alpha)}]. \tag{2.2}$$

When going from (2.1) to (2.2) we use $\hat{\theta} - \theta$ as a pivotal quantity. This concept has been developed further by Beran (1987). In some situations there might exist a transformation f such that $f(\hat{\theta}) - f(\theta)$ is more pivotal than $\hat{\theta} - \theta$. The basic confidence interval method is then applied to $f(\theta)$, and the confidence interval for θ is obtained by transforming the confidence

[1] The * indicates that the variables and estimators are simulated, and later on that probabilities, expectations and variances are with respect to the sampling distribution.

limits for $f(\theta)$ back to the original scale using the inverse of f. Such intervals will be referred to as *basic f.*

A different approach, which seems to first appear in Efron (1981b), is based on the existence of a transformation h such that $h(\hat{\theta}) - h(\theta)$ and $h(\hat{\theta}^*) - h(\hat{\theta})$ have the same symmetric distribution around 0 for all values of θ and $\hat{\theta}$. Under these assumptions it can be shown that the *percentile* interval

$$[\hat{\theta}^*_{(\alpha)}, \hat{\theta}^*_{(1-\alpha)}], \tag{2.3}$$

which is read off directly from the bootstrap distribution of $\hat{\theta}^*$, has coverage probability $1-2\alpha$. Although the percentile interval seems less appealing from a theoretical point of view, if the parameter space is finite and the estimator is confined to that space, the percentile interval will always produce confidence intervals within that space, but this will not always be the case with the basic interval.

Two methods of calculating confidence intervals, related to the percentile interval, are the *bias corrected percentile* method, BC, and the *accelerated bias corrected percentile* method, BCa. The assumption of the latter is that there exist a monotone increasing transformation h, constants a and z_0, such that

$$\frac{h(\hat{\theta}) - h(\theta)}{1 + ah(\theta)} + z_0, \tag{2.4}$$

and its bootstrap version, is $N(0,1)$ for all θ. The BC method uses $a = 0$, and with both a and z_0 equal to zero we get the percentile interval. From the bootstrap we get the distribution of $\hat{\theta}^*$, but under the assumption that the bootstrap version of (2.4) is $N(0,1)$ we also have

$$\alpha = \mathsf{P}^* \left(\hat{\theta}^* \leq h^{-1} \left[h(\hat{\theta}) + (z^{(\alpha)} - z_0)(1 + ah(\hat{\theta})) \right] \right),$$

where $z^{(\alpha)}$ is the α-percentile of the $N(0,1)$-distribution. Thus, we have

$$h^{-1} \left(h(\hat{\theta}) + (z^{(\alpha)} - z_0)(1 + ah(\hat{\theta})) \right) = \hat{\theta}^*_{(\alpha)}. \tag{2.5}$$

A $1-\alpha$ upper confidence limit for θ is now obtained by using (2.4) and (2.5):

$$\begin{aligned} 1-\alpha &= \mathsf{P}\left(z^{(\alpha)} < \tfrac{h(\hat{\theta})-h(\theta)}{1+ah(\theta)} + z_0\right) \\ &= \mathsf{P}\left(\theta < h^{-1}\left[h(\hat{\theta}) + \tfrac{z_0-z^{(\alpha)}}{1-a(z_0-z^{(\alpha)})}(1+ah(\hat{\theta}))\right]\right) \\ &= \mathsf{P}\left(\theta < h^{-1}\left[h(\hat{\theta}) + (z^{(\beta_2)}-z_0)(1+ah(\hat{\theta}))\right]\right) \\ &= \mathsf{P}\left(\theta < \hat{\theta}^*_{(\beta_2)}\right) \end{aligned} \tag{2.6}$$

where $\beta_2 = \beta(1-\alpha)$ and

$$\beta(\alpha) = \Phi(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}).$$

are obtained by equating the mid-rows of (2.6). Thus, with $\beta_1 = \beta(\alpha)$ we get the $1-2\alpha$ two-sided confidence BCa interval for θ :

$$[\hat{\theta}^*_{(\beta_1)}, \hat{\theta}^*_{(\beta_2)}]. \tag{2.7}$$

However, one thing remains to be done, and that is to set the bias parameter z_0 and the acceleration parameter a. Starting with the easy one of the two, the bias parameter, by using (2.4) and the fact that h is monotone increasing we get $\mathsf{P}^*(\hat{\theta}^* \le \hat{\theta}) = \Phi(z_0)$, where $\mathsf{P}^*(\hat{\theta}^* \le \hat{\theta})$ is the proportion of $\hat{\theta}^*$ in the bootstrap that fall below $\hat{\theta}$. Thus we get the bias parameter as

$$z_0 = \Phi^{-1}\left(\mathsf{P}^*(\hat{\theta}^* \le \hat{\theta})\right). \tag{2.8}$$

The acceleration parameter a is not as easy to find, but a non-parametric approximation was suggested by Efron (1987),

$$a = \frac{\sum_{i=1}^n \psi_i^3}{6(\sum_{i=1}^n \psi_i^2)^{3/2}}, \tag{2.9}$$

where

$$\psi_i = \lim_{\Delta\to 0} \frac{\hat{\theta}((1-\Delta)F_n + \Delta\xi_i) - \hat{\theta}(F_n)}{\Delta}$$

is an empirical influence measure of observation x_i on $\hat{\theta}$. Here $\hat{\theta}$ is treated as a function of distributions, and ξ_i is the degenerated distribution with a unit mass at x_i.

2.2 Resampling cases, binomial model

The "classical" way of performing bootstrap is to resample from the observations with replacement and then rely on that the distribution of the bootstrap sample will mimic the distribution of the original sample. Here we will compare six different ways of forming confidence intervals that are based on this "classical" way: percentile, basic, basic with logit and arcsine transformation, BC and BCa.

2.2.1 Bootstrap distribution

Let x be the observed number of "successes" out of n independent Bernoulli variables, all with the same unknown "success"-probability p, i.e., x is the observed value of a binomial variable with parameters n and p. The obvious estimator of p based on these observations is $\hat{p} = x/n$.

When we perform the bootstrap we draw from the n "successes" and "failures" with replacement, thus we have probability x/n of getting a "success" in each of the n draws. If we let x^* be the number of "successes" in a bootstrap sample, then x^* is an observation of a binomial variable with parameters n and x/n.

The different ways of forming confidence intervals that we will compare below are all computed using the inverse of the empirical distribution function. But since the bootstrap distribution is known, we also know what values the empirical distribution function will converge[2] to as the number of bootstrap samples tends to infinity. Thus, we can compare the intervals themselves and their coverage probabilities without any tedious simulations.

2.2.2 Confidence intervals

In figure 2.1 the confidence limits for every possible relative frequency $\hat{p}$ and the coverage probabilities of the lower limit for $0 \leq p \leq 1$ are displayed, $n = 30$. The coverage probability of the upper limit at p equals the coverage probability of the lower limit at $1 - p$ so we need only study one of them. The upper graphs contain the confidence limits of the different bootstrap intervals (solid line) together with the exact two-sided intervals (dashed line) for the parameter p. The intervals in the figure are from left to right:

[2] almost surely

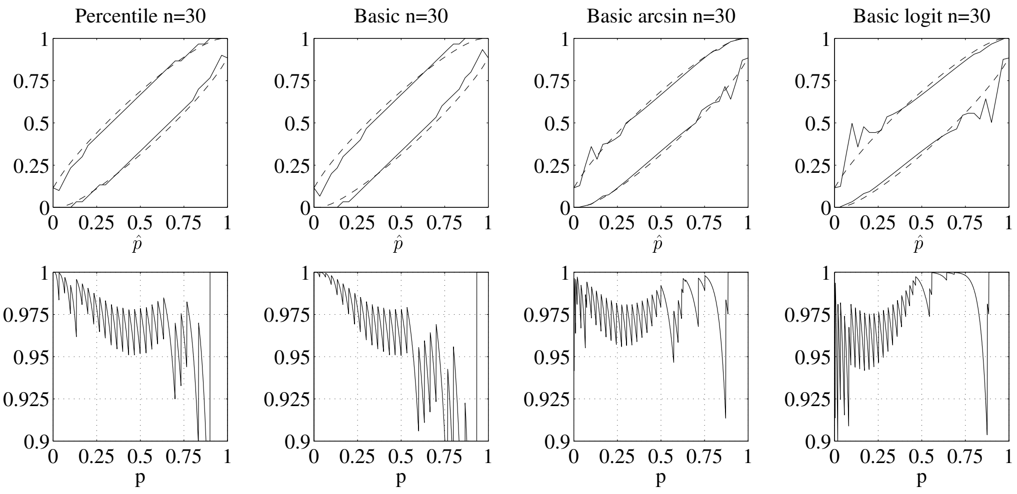

Figure 2.1: Two-sided 95% confidence intervals (upper graphs) and the coverage probabilities of the lower limits (lower graphs).

Percentile	:	$[\hat{p}_\alpha, \hat{p}_{1-\alpha}]$
Basic	:	$[2\hat{p} - \hat{p}_{1-\alpha}, 2\hat{p} - \hat{p}_\alpha]$,
Basic with transformation f	:	$[\mathrm{f}^{-1}(2\mathrm{f}(\hat{p}) - \mathrm{f}(\hat{p}_{1-\alpha})), \mathrm{f}^{-1}(2\mathrm{f}(\hat{p}) - \mathrm{f}(\hat{p}_\alpha))]$

where f is either $\arcsin(\sqrt{\cdot})$ or a modified version of the logit function [3] .

When $\hat{p}$ is 0 or 1 the bootstrap gives $\hat{p}_\alpha = \hat{p} = \hat{p}_{1-\alpha}$. Thus all four intervals will give the confidence interval $[0,0]$ or $[1,1]$ respectively. These are obviously not very good intervals, and therefore the confidence limits are replaced with the exact limits. This kind of modification may seem strange and ad-hoc to do, but to prevent the limits corresponding to $\hat{p}$ being 0 or 1 from disturbing the coverage probability comparisons, we can not use the intervals as they are. Another way around this problem could be to consider the coverage probability conditionally on $0 < \hat{p} < 1$, but then it would be harder to interpret the result for p close to 0 or 1.

Both the percentile and the basic intervals show a "bias", compared to the exact interval, in the sense that both limits are either smaller or larger than the corresponding exact limits. This "bias" is most pronounced for the basic interval. Moreover, the coverage probability of the lower limit is far too low for large p. This is true for both intervals, but especially for the basic

[3] $\mathrm{logit}_{modified}(p) = \log(p + \delta) - \log(1 + \delta - p)$, with $\delta = 1/2n$

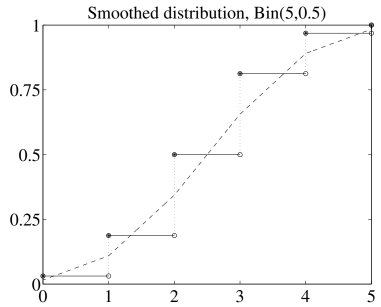

Figure 2.2: Smoothed distribution.

interval. For both intervals the upper limit for $\hat{p}$ equals $1/n$ is smaller than the limit for $\hat{p} = 0$, but this inconsistency is caused by the replacement of the upper limit with the exact one at $\hat{p} = 0$.

The basic arcsine and the basic logit intervals look a bit strange, especially the basic logit interval. At $\hat{p} = 0.1$ the upper limit of the basic arcsine interval is bigger than the upper limit at $\hat{p} = 0.1333$. But otherwise the basic arcsine interval is closest to the exact interval overall. The basic logit interval has two points where the upper limit decreases in $\hat{p}$. Also when looking at the coverage probabilities the basic arcsine interval performs better than the other three, but on the other hand, none of the four confidence intervals performs satisfactory.

2.2.3 Smoothed bootstrap distribution

When we calculate the percentiles of the bootstrap distribution as described above we use the distribution function of $\hat{p}^*$ as it is. One possibility is to use a smoothed and continuous version instead. A very simple way of smoothing is to use a linear spline as described in figure 2.2.

The effects of using the smoothed distribution instead of the empirical distribution on the percentile and basic arcsine intervals are found in figure 2.3. For the percentile interval there is very little improvement. Looking at the basic arcsine interval we see that the decrease in the upper limit is gone, but we still have a bump on the lower limits around $\hat{p} = 0.1$ and the lower limit is too big for $\hat{p}$ above 0.95. However, besides this, the deviance from the exact interval is very small and the coverage probability of the lower limit for $p < 0.5$ is almost perfect. But, on the other hand, the coverage

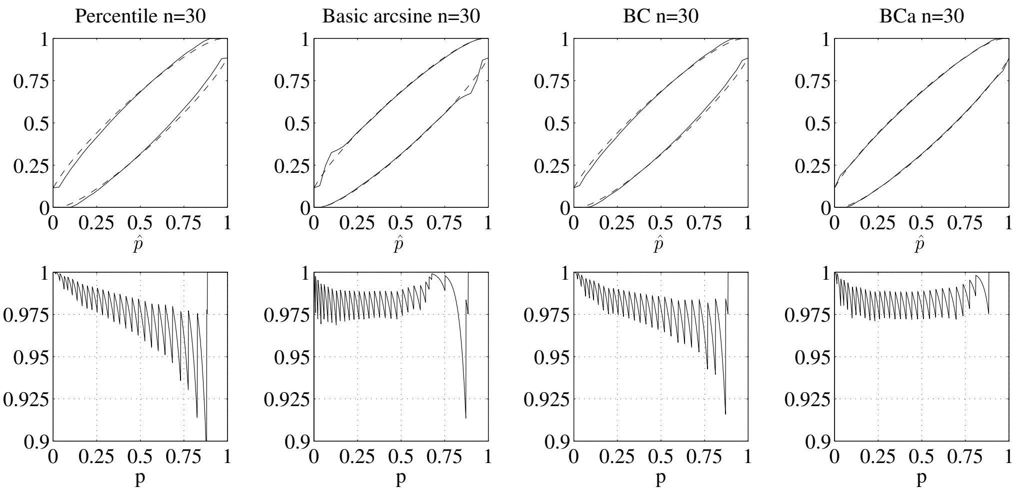

Figure 2.3: Two-sided 95% confidence intervals (upper graphs) and the coverage probabilities of the lower limits (lower graphs), smoothed versions.

probability drops well below the target 0.975 for p around 0.85, and the bump causes a big conservatism around p equal to 0.7.

In the same figure two other intervals are also shown, BC and BCa. The BCa interval involves a bias parameter, z_0, and an acceleration parameter, a, as described on page 5, whereas the BC interval only involves the bias parameter. Instead of using the z_0 in (2.8) we will use

$$z_0 = \Phi^{-1}\left(\mathsf{P}^*(\hat{\theta}^* < \hat{\theta}) + 0.5\mathsf{P}^*(\hat{\theta}^* = \hat{\theta})\right).$$

With this modification the intervals of θ and $1-\theta$ will not contradict one-another. Further, ψ_i, the empirical influence measure of observation i on $\hat{\theta}$, is $-\hat{\theta}$ for all failures, and $1-\hat{\theta}$ for all successes. With these empirical influence measures we get the acceleration parameter of (2.9):

$$a = \frac{1-2\hat{\theta}}{6\sqrt{n\hat{\theta}(1-\hat{\theta})}}.$$

In the figure we see that the performance of the BC interval is only slightly better than the percentile interval, but the coverage is still not as good as for the smoothed basic arcsine interval. But looking at the BCa interval and

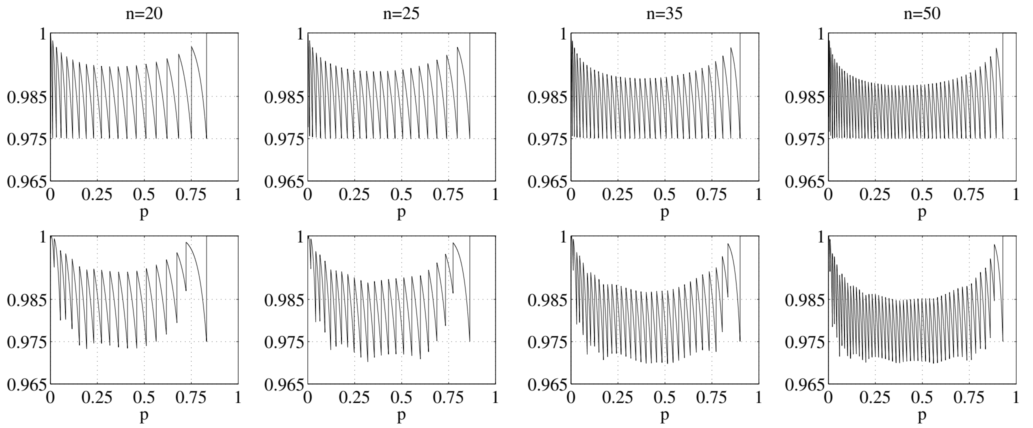

Figure 2.4: Coverage probabilities of the 97.5% lower limit, exact limits (upper graphs) and BCa limits (lower graphs).

its coverage probability, the figure shows a dramatic improvement compared to the percentile interval. Of the intervals tested here BCa is the only one which does not show a coverage probability well below the intended level at any point.

The coverage probability of the exact interval and the BCa interval of four different values of n is found in figure 2.4. Compared to the exact interval, the BCa interval is more conservative when p is either quite small or large, while for p around 0.5 the coverage probability range of the BCa interval falls below 0.975, but in these four examples never below 0.97, which seems quite acceptable.

2.3 Exact confidence intervals

As done in the previous section, exact confidence intervals for binomial data can be calculated quite easily, see e.g. (Collett, 1991). Suppose x is an observation of a binomial(n,p) distributed variable X. A $1-\alpha$ upper bound for p is obtained by testing the hypothesis $H_{p_0} : p \geq p_0$ for all possible values of p_0. The hypothesis H_{p_0} is rejected when $\mathsf{P}_{p_0}(X \leq x) \leq \alpha$. Since $\mathsf{P}_{p_0}(X \leq x)$ decreases with p_0 the upper bound for p is the value of p_0 that solves the equation $\mathsf{P}_{p_0}(X \leq x) = \alpha$. If there is no solution to the equation, i.e. if $\mathsf{P}_{p_0}(X \leq x) > \alpha$ for all p_0, the upper bound for p is 1. The upper

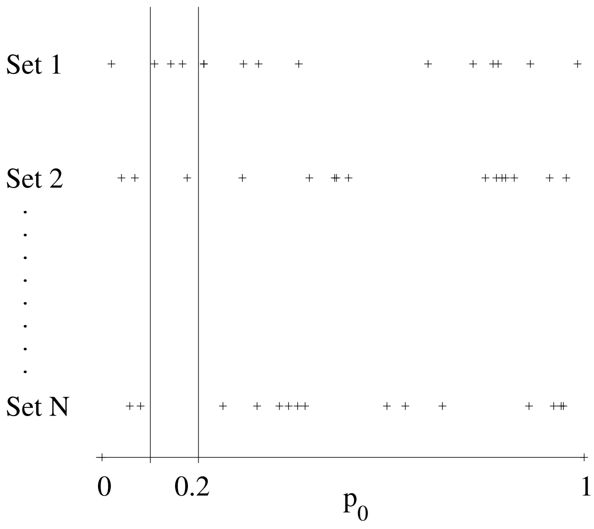

Figure 2.5: Calculating exact confidens intervals for binomial data. Example with n equal to 15. The $+$'s are the simulated uniformly distributed variables.

bound can be obtained by numerically solving the equation above.

However, it is also possible to do the same thing by means of simulations, in a rather elegant way. Step one is to simulate N sets, each sized n, of independent variables uniformly distributed on the interval $(0, 1)$. For each set, $x^*(p_0)$ is the number of simulated variables less than or equal to p_0. Then each $x^*(p_0)$ is an observation of a binomial(n,p_0) distributed variable. The result of the simulation could look like figure 2.5. In the first set $x^*(0.1)$ is equal to 1, and $x^*(0.2)$ equals 4. In set 2 $x^*(0.2)$ is equal to 3.

The relative frequency of sets satisfying $x^*(p_0) \leq x$ is an estimate of $\mathsf{P}_{p_0}(X \leq x)$, and the upper bound for p is the value of p_0 for which $\alpha \cdot N$ of the sets fulfil $x^*(p_0) \leq x$. Since each $x^*(p_0)$ is a non-decreasing and right-continuous function of p_0, the number of sets fulfilling $x^*(p_0) \leq x$ is non-increasing. Hence, we find the estimate of the upper bound in the following way: For each set, let p^* be the smallest value of p_0 for which $x^*(p_0)$ is greater than x:

$$p^* = \min\{p_0 : x^*(p_0) > x\} = \sup\{p_0 : x^*(p_0) \leq x\}.$$

Let $p^*_{(1-\alpha)}$ be the $(1-\alpha) \cdot N$'th of the ordered p^*. Then, assuming there are no ties among the p^*'s, all $\alpha \cdot N$ sets with a $p^* > p^*_{(1-\alpha)}$ fulfil $x^*(p^*_{(1-\alpha)}) \leq x$,

N	100	200	500	1 000	5 000	10^4	10^5
$p^*_{(1-\alpha)}$	0.9403	0.9506	0.9492	0.9424	0.9457	0.9440	0.9434

Table 2.1: Exact upper bound calculated by means of simulation.

while none of the $(1-\alpha)\cdot N$ sets with $p^* \leq p^*_{(1-\alpha)}$ satisfies the same criterion. Thus, the estimate of the upper bound is $p^*_{(1-\alpha)}$.

Let $u^*_{(1)}, u^*_{(2)}, \ldots, u^*_{(n)}$ denote the ordered simulated variables of a set. If x is less than n, the definition of p^* implies $p^* = u^*_{(x+1)}$. So what it boils down to, is that the upper bound is just the $1-\alpha$ percentile of the distribution of an order statistic, which we compute by means of simulation. An example with $n = 15$, $x = 12$ and N ranging from 100 up to 100 000 is found in table 2.1. The correct upper bound is 0.9432.

2.4 Overdispersed binomial model

In this section we will continue with Bernoulli variables of which we want to estimate the unknown success probability θ. But now we will study a hierarchical model by introducing a distribution P for the success parameter.

In detail: assume that we have m groups, each sized n. The number of successes in group i, x_i, conditional on p_i, is binomial distributed with parameters n and p_i. The parameters $p_1, \ldots, p_m$ are independent random variables distributed according to the distribution P, with mean θ and variance $\nu\theta(1-\theta)$, $0 \leq \nu \leq 1$.

If x_{ij} is the success indicator for observation j of group i we have $\mathrm{Var}(x_{ij}) = \theta(1-\theta)$ and $\mathrm{Cov}(x_{ij}, x_{il}) = \nu\theta(1-\theta)$ for $j \neq l$, while the covariance between x_{ij} and x_{kl} is zero when $i \neq k$. Our estimate of θ is the total relative frequency of successes, $\hat{\theta} = \sum_{i=1}^{m} x_i/(mn)$ which has variance

$$\mathrm{Var}\left(\hat{\theta}\right) = \frac{\theta(1-\theta)}{mn}\left[1 + (n-1)\nu\right]. \tag{2.10}$$

In Davison and Hinkley (1997) two simple methods of bootstrapping hierarchical data are studied. Both are done in two steps. In the first step each method sample m groups with replacement. In the second step observations of each group sampled in step 1 are sampled. Suppose group 1 was selected once in step 1. Then method 1 samples among the n observations of group 1 without replacement, while method 2 samples n observations with replacement.

2.4.1 Bootstrap method 1

When bootstrapping according to method 1 we actually only sample groups. If group i_0 is sampled k times in step 1, all observations of group i_0 are included exactly k times in the whole bootstrap sample. We can think of $\hat{\theta}^*$ as an average of m observations from the uniform distribution on $\{\hat{p}_1, \dots \hat{p}_m\}$, where $\hat{p}_i = x_i/n$. Thus

$$\mathrm{Var}_1^*\left(\hat{\theta}^*\right) = \frac{1}{m}\mathrm{Var}^*(\hat{p}^*) = \frac{1}{m^2}\sum_{i=1}^{m}\left(\hat{p}_i - \hat{\theta}\right)^2 = \frac{1}{m^2}\sum_{i=1}^{m}\hat{p}_i^2 - \frac{1}{m}\hat{\theta}^2.$$

Calculating the expected variance we get

$$\mathsf{E}\left[\mathrm{Var}_1^*\left(\hat{\theta}^*\right)\right] = \frac{\theta(1-\theta)}{mn}\left[1 + (n-1)\nu\right]\frac{m-1}{m}. \qquad (2.11)$$

Although the model of the bootstrap sample does not have the same structure as the model of the original data the expected variance is very close to the variance of $\hat{\theta}$ as long as m is large enough. Regardless of the value of n, it is asymptotically correct.

2.4.2 Bootstrap method 2

The bootstrap sample obtained when bootstrapping according to method 2 will have the same structure as the original data. In the first step we sample with replacement from all the groups, suppose i^* is the index of the sampled group. Then we sample n times with replacement from the x_{i^*} successes and $n - x_{i^*}$ failures with replacement. Thus conditional on the sampled group, the number of successes of that group is binomial distributed with parameters n and $\hat{p}_{i^*} = x_{i^*}/n$. The only difference of the bootstrap model compared to the model of the original data is the distribution of the p^*'s which in the bootstrap model is uniform on $\{\hat{p}_1, \dots, \hat{p}_m\}$, thus $\mathsf{E}^*[p^*] = \hat{\theta}$ and $\mathrm{Var}^*(p^*) = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{p}_i - \hat{\theta}\right)^2$. This yields

$$\mathrm{Var}_2^*\left(\hat{\theta}^*\right) = \frac{\hat{\theta}}{mn} - \frac{\hat{\theta}^2}{m} + \frac{n-1}{m^2 n}\sum_{i=1}^{m}\hat{p}_i^2$$

and

$$\mathsf{E}\left[\mathrm{Var}_2^*\left(\hat{\theta}^*\right)\right] = \frac{\theta(1-\theta)}{mn}\left[\left(2-\frac{1}{m}-\frac{1}{n}\right)+(n-1)\nu\left(1-\frac{1}{m}-\frac{1}{n}\right)\right]. \quad (2.12)$$

When using method 2 the expected variance of the bootstrap estimate of θ can deviate quite much from the variance of $\hat{\theta}$. If in fact all observations where independent, i.e., P is degenerated with point-mass at θ and $\nu = 0$, the bootstrap observations will still have a within-group dependence unless all $\hat{p}_i$ are equal. Only when $n = 1$ the expected variance of the bootstrap estimate of θ is asymptotically correct.

2.4.3 Bootstrap method 3

A third simple way of bootstrapping hierarchical data is of course to neglect the dependence and bootstrap as if all observations were independent. In our situation this results in $mn\hat{\theta}^*$ being binomial distributed with parameters mn and $\hat{\theta}$ and the variance of $\hat{\theta}^*$ to be

$$\mathrm{Var}_3^*\left(\hat{\theta}^*\right) = \frac{\hat{\theta}(1-\hat{\theta})}{mn}$$

and

$$\mathsf{E}\left[\mathrm{Var}_3^*\left(\hat{\theta}^*\right)\right] = \frac{\theta(1-\theta)}{mn}\left[\left(1-\frac{1}{mn}\right)+(n-1)\nu\frac{1}{mn}\right]. \quad (2.13)$$

If ν is zero this way of bootstrapping will of course work just fine, but as ν gets larger the more will the variance on average be underestimated.

2.4.4 Which one?

When using method 1 we do not get the same structure of model but the expected variance is close to the desired one. With method 2 we get the same structure of the model but the expected variance can be as much as two times bigger than the desired one. With method 2 we will always have a bigger variance than with method 1.

In figure 2.6 the ratio of expected variance of method 2 and 3 to the true variance are shown. The maximal ratio of method 2, maximum over n and m, in the left graph declines quite rapidly as ν gets bigger. If, for instance, ν is at least $1/4$ the ratio is never greater than $4/3$, and for ν above 0.5 the worst case is 1.18. On the other hand, as the dependence gets larger, the more

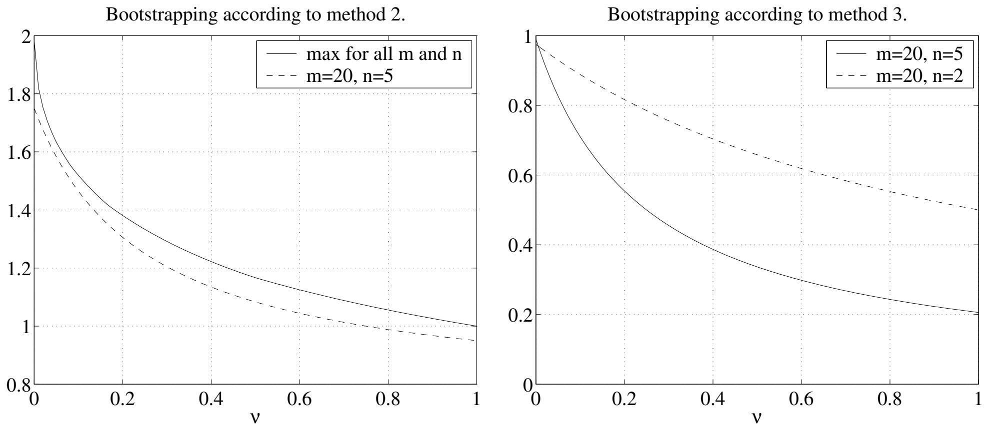

Figure 2.6: Ratio of the expected variance of method 2 and 3 to the true variance.

will method 2 resemble method 1. When bootstrapping as if all observations were independent there is no lower limit for how much the variance will be underestimated.

2.5 Summary

In section 2.2 six different ways of using bootstrap to compute confidence intervals were compared. They all were based on bootstrap samples obtained by sampling cases. The BCa interval, when using a smoothed bootstrap distribution, produced coverage probabilities not far from the exact interval. All other intervals had ranges for p for which the coverage probability was too low.

We have also seen a way of computing the exact confidence interval by using bootstrap, where the same bootstrap can represent samples from different models.

When all Bernoulli variables no longer are independent, specifically, when groups of variables have positive correlation, the robust way of bootstrapping is to sample groups only. The alternative method of sampling in two steps, first group and the within group, will produce bootstrap distribution which are too wide. Sampling as if all observations are independent will produce too narrow bootstrap distributions.

Chapter 3

Univariate survival data

In the random censoring model there are two survival functions. The failure time, Y^0, and the censoring time, C, are independent and distributed according to survival functions F and G respectively. We observe only the minimum of Y^0 and C, and an indicator of which one is the smallest. The survival function of interest is F, whereas G is only a nuisance parameter. Suppose we want to estimate the survival rate at t_0, $\theta = F(t_0)$. For this purpose we use is the KM evaluated at t_0, $\hat{\theta} = \hat{F}(t_0)$.

In this chapter we will consider two bootstrap techniques for calculating confidence intervals for θ. First we will introduce a technique for calculating confidence intervals that resemble the exact confidence intervals for the binomial model, computed using bootstrap. The second bootstrap technique we will study is the accelerated bias corrected percentile method, BCa.

3.1 Semiparametric exact confidence interval

Suppose we have n independent observations of $Y = \min(Y^0, C)$ and $\delta = I(Y^0 \leq C)$, $\{(y_i, \delta_i), i = 1, \ldots, n\}$. Let there be k distinct times of failures and censorings, $\tau_1, \ldots, \tau_k$. If there are ties between failures and censored observations, the censorings are taken to occur immediately after the failures, and correspond to different τ's. Let $r_j = \sum_i I(y_i \geq \tau_j)$ be the number of observations at risk at τ_j, d_j and c_j are the number of failures and censored observations at τ_j respectively. Note that since we separate ties among failures and censorings $d_j \cdot c_j = 0$ for all j.

From these observations we compute KM estimates of both F and G:

$$\begin{aligned}\hat{F}(t) &= \prod_{\tau_j \leq t} (1 - \frac{d_j}{r_j}) \\ \hat{G}(t) &= \prod_{\tau_j \leq t} (1 - \frac{c_j}{r_j}).\end{aligned}$$

To draw with replacement from the pairs $\{(y_i, \delta_i), i = 1, \ldots, n\}$, is the same as simulating from the random censoring model with survival functions $\hat{F}$ and $\hat{G}$ (Davison & Hinkley, 1997). On such a bootstrap sample we can calculate the KM estimator, $\hat{F}^*$, and evaluate it at t_0 giving $\hat{\theta}^* = \hat{F}^*(t_0)$. By such a simulation we will get a picture of the distribution of $\hat{\theta} = \hat{F}(t_0)$ under the model where $F = \hat{F}$ and $G = \hat{G}$. But in order to compute intervals we want to know the distribution of $\hat{\theta}$ under models where $F(t_0)$ has other values than the observed $\hat{\theta}$, and find values which make our estimate $\hat{\theta}$ unlikely.

3.1.1 Choice of alternative model

A fully non-parametric confidence interval for θ would require that we knew the distribution of $\hat{\theta}$ under all pairs of survival functions F and G. But since this family is far too big to work with we will only consider pairs of F and G such that the F's are some family of stochastically ordered distributions generated from $\hat{F}$, and G is held fixed at $\hat{G}$. The question is how to choose the family of stochastically ordered distributions.

Since $\hat{F}$ is an unconstrained ML estimator of F a natural choice seems to be using a constrained ML estimator of F. The constrained estimator is easily obtained by introducing a Lagrangian multiplier to the log likelihood function (Thomas & Grunkemeier, 1975). In detail, using the constraint $\hat{F}(t_0) = \theta_0$ we get

$$\hat{F}_\lambda(t) = \prod_{j:\tau_j \leq \min(t_0, t)} (1 - \frac{d_j}{r_j + \lambda}) \prod_{j:t_0 < \tau_j \leq t} (1 - \frac{d_j}{r_j}), \tag{3.1}$$

where $\lambda = \lambda(\theta_0)$ satisfies $\hat{F}_\lambda(t_0) = \theta_0$. Since $\hat{F}_\lambda(t)$ is non-decreasing in λ for every t, the obtained family is stochastically ordered.

By means of simulation we can estimate the distribution of $\hat{\theta}$ under $(\hat{F}_\lambda, \hat{G})$ for any $\lambda > d_j - r_j$, where τ_j is largest distinct failure time less

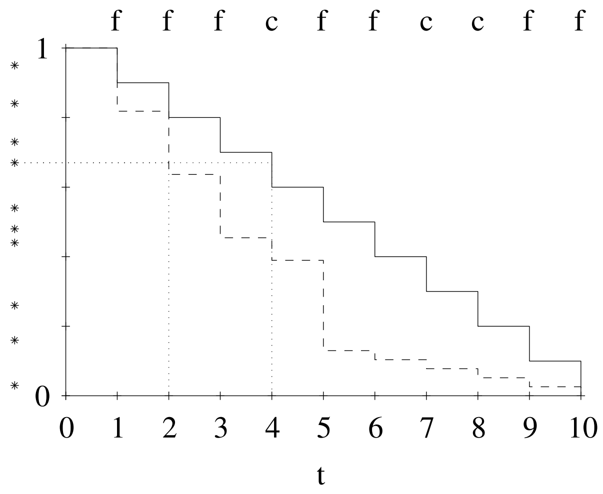

Figure 3.1: Simulation of failure/censor times, Y^*'s under different models, $(\hat{F}_0, \hat{G})$ (solid line) and $(\hat{F}_{-4.5}, \hat{G})$ (dashed line).

than or equal to t_0. We could do this by simulating $Y^{0*} \sim \hat{F}_\lambda$ and $C^* \sim \hat{G}$ and then calculate (Y^*, δ^*) as before. But we might just as well simulate Y^* directly, using $\hat{F}_\lambda(t) \cdot \hat{G}(t)$ and just keep track of the original observation being censored or not. This corresponds to weighted sampling from the pairs $\{(y_i, \delta_i), i = 1, \dots, n\}$.

3.1.2 An example

Figure 3.1 shows an example on how these simulations can be done. The data consists of 10 observations of which 3 are censored. The observations are displayed over the graph, f for failure and c for a censored observation. Suppose we want to find a confidence interval for F at $t_0 = 5$. The KM estimator at this point equals 0.5833 (7/12).

The $*$'s beside the vertical axis are 10 simulated U(0,1) distributed variables, $u_1^*, \dots, u_{10}^*$, which we use to simulate y^*. The figure shows how we can use the same set of u^*-variables for different models. As seen by the dotted lines, when using $(\hat{F}, \hat{G})$ (solid line), which corresponds to random draw with replacement from the 10 observations, the u^* equal to 0.67 will produce $y^* = 4$ and $\delta^* = 0$. But when using $(\hat{F}_{-4.5}, \hat{G})$ it will produce $y^* = 2$

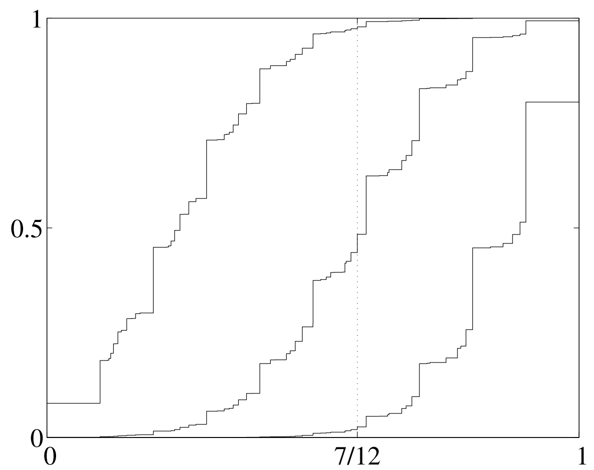

Figure 3.2: Distribution of $\hat{\theta}^*$ under $(\hat{F}_\lambda, \hat{G})$ for λ equal to -3.975, 0.0, and 16.25.

and $\delta^* = 1$. The complete bootstrap samples are $\{1, 2, 3, 4^c, 5, 6, 6, 8^c, 9, 10\}$ when using $(\hat{F}, \hat{G})$, and $\{1, 1, 2, 2, 3, 3, 4^c, 5, 5, 9\}$ using $(\hat{F}_{-4.5}, \hat{G})$.

Figure 3.2 shows the result from simulations under three different values on λ. Each curve is based on 100 000 simulations. The left, middle and right curve is the estimated distribution of $\hat{\theta}^*$ under $(\hat{F}_\lambda, \hat{G})$ for λ equal to -3.975, 0.0 and 16.25 respectively. From these curves we obtain the estimated probabilities and expectations of table 3.1. From the table we conclude that a $1 - 0.0251 - 0.0254 \approx 0.95$ confidence interval for θ would be

$$[0.2541, 0.8459].$$

3.1.3 Special cases

If $\hat{\theta}$ equals 0 or 1 we actually do not need to use bootstrap to compute the exact upper and lower limits respectively. Instead we can find the limits by numerically computing some probabilities under the constrained ML estimator $\hat{F}_\lambda$ and $\hat{G}$. The lower and upper limits are naturally 0 and 1 respectively. In this section variables marked with a * are distributed according to $\hat{F}_\lambda$ or $\hat{G}$.

λ	$\hat{F}_\lambda(5)$	$\mathsf{P}_{(\hat{F}_\lambda,\hat{G})}\left(\hat{\theta}^* \leq \hat{\theta}\right)$	$\mathsf{P}_{(\hat{F}_\lambda,\hat{G})}\left(\hat{\theta}^* \geq \hat{\theta}\right)$	$\mathsf{E}_{(\hat{F}_\lambda,\hat{G})}\left[\hat{\theta}^*\right]$
-3.975	0.2541	0.9795	0.0251	0.2547
0	0.5833	0.4848	0.5587	0.5831
16.25	0.8459	0.0254	0.9811	0.8456

Table 3.1: Probabilities and expectations under different models.

$\hat{\theta}$ equals 0

When $\hat{\theta}$ is equal to 0 we have observations such that the largest one, $Y_{(n)}$, is less than t_0. We also know that this was a failure, i.e. the corresponding failure indicator $\delta_{(n)}$, is equal to 1. To find the upper confidence limit we need to calculate the probability of the event $\hat{\theta}^* = 0$. But this event is equivalent with $(Y^*_{(n)} \leq t_0) \cap (\delta^*_{(n)})$, and we can find the probability of the latter by conditioning on the value of $Y^*_{(n)}$:

$$\begin{aligned}
\mathsf{P}_\lambda\left((Y^*_{(n)} \leq t_0) \cap (\delta^*_{(n)} = 1)\right) &= \sum_{j=1}^{k} \mathsf{P}_\lambda\left(Y^*_{(n)} = \tau_j\right) \cdot \\
&\qquad \mathsf{P}_\lambda\left((Y^*_{(n)} \leq t_0) \cap (\delta^*_{(n)} = 1) | Y^*_{(n)} = \tau_j\right) \\
&= \sum_{j=1}^{k} \mathsf{P}_\lambda\left(Y^*_{(n)} = \tau_j\right) \mathrm{I}(d_j > 1).
\end{aligned}$$

Further, since

$$\mathsf{P}_\lambda\left(Y^*_{(n)} \leq \tau_j\right) = \left(\mathsf{P}_\lambda\left(Y^*_1 \leq \tau_j\right)\right)^n = \left(1 - \hat{F}_\lambda(\tau_j)\hat{G}(\tau_j)\right)^n$$

and both $\hat{F}_\lambda$ and $\hat{G}$ are known functions, we can numerically find the probability of $\hat{\theta}^* = 0$. The upper confidence limit is then obtained by numerically solving the equation $\mathsf{P}_\lambda\left(\hat{\theta}^* = 0\right) = \alpha/2$ with respect to λ.

$\hat{\theta}$ equals 1

In this case the original sample has no failures before t_0. All observations less than t_0, if any, are censored ones. Suppose there are s censored observations less than t_0. What we need to calculate now is the probability of the event

$\hat{\theta}^* = 1$, which equals

$$\mathsf{P}_\lambda\left(\hat{\theta}^* = 1\right) = \mathsf{P}_\lambda\left(\bigcap_{i=1}^{n}\left(Y_i^{0*} > C_i^* \wedge t_0\right)\right) = \left(\mathsf{P}_\lambda\left(Y_1^{0*} > C_1^* \wedge t_0\right)\right)^n.$$

This probability is computed under the constrained ML estimator $\hat{F}_\lambda$ and $\hat{G}$. However, when there are no failures in the interval $[0, t_0]$ the constrained ML estimator of F is not unique. The probability $1 - \theta_0$ can be arbitrarily distributed over the interval $[t', t_0]$ where t' is the largest censored observation less than t_0, if there are any, zero otherwise. By putting the point mass $1 - \theta_0$ at t_0 we get

$$\begin{aligned}
\mathsf{P}_\lambda\left(Y_1^{0*} > C_1^* \wedge t_0\right) &= \sum_{j:\tau_j < t_0} \mathsf{P}_\lambda\left(C_1^* = \tau_j\right)\mathsf{P}_\lambda\left(Y_1^{0*} > C_1^* \wedge t_0 | C_1^* = \tau_j\right) + \\
&\qquad \mathsf{P}_\lambda\left(C_1^* \geq t_0\right)\mathsf{P}_\lambda\left(Y_1^{0*} > C_1^* \wedge t_0 | C_1^* \geq t_0\right) \\
&= \mathsf{P}_\lambda\left(C_1^* < t_0\right) + \mathsf{P}_\lambda\left(C_1^* \geq t_0\right)\mathsf{P}_\lambda\left(Y_1^{0*} > t_0\right)
\end{aligned}$$

since under $\hat{F}_\lambda$ we have $Y_1 \geq t_0$ with probability 1. Moreover, $\mathsf{P}_\lambda\left(Y_1^{0*} > t_0\right)$ is just θ_0 and because $\mathsf{P}_\lambda\left(C_1^* \geq t_0\right) = (n - s)/n := P_g$ we end up with $\mathsf{P}_\lambda\left(\hat{\theta}^* = 1\right) = (1 - P_g + P_g\theta_0)^n$. Finally, setting this equal to $\alpha/2$ yields the lower confidence limit

$$\frac{\exp\{\log(\alpha/2)/n\} - 1 + P_g}{P_g}.$$

3.1.4 Computational drawback

However, there is one major drawback with this technique: Holding the set of u^*'s fix, and varying λ, i.e. varying the model, the KM estimator of the bootstrap sample at t_0, $\hat{\theta}^*$, is not a monotone function of λ. When we decrease λ we get a model which is stochastically smaller. It would then be desirable that the KM estimator of the bootstrap sample decreased, or at least did not increase. But $\hat{\theta}^*$ can both increase and decrease when we decrease λ.

What happens as we change λ, is that the value of one or several bootstrap observations change. If we decrease λ, the bootstrap observations will become smaller. But the bootstrap observations can also change from being a failure to become a censored observation or vice-verse, and this is what causes the

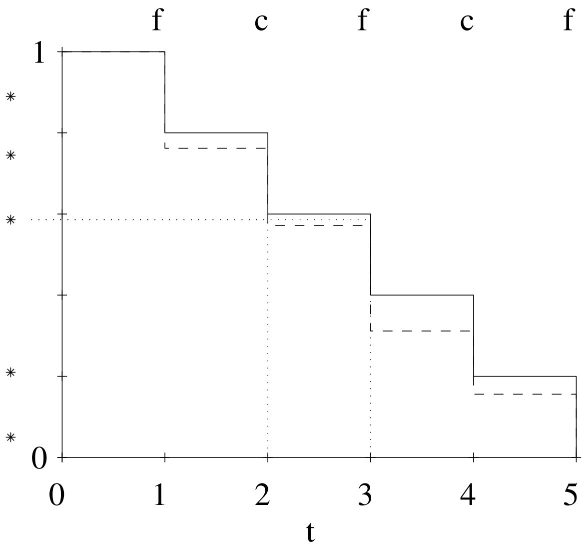

Figure 3.3: Simulation of failure/censor times, Y^*'s under different models, $(\hat{F}_0, \hat{G})$ (solid line) and $(\hat{F}_{-0.8}, \hat{G})$ (dashed line).

non-monotonicity. If we would sample both y^{0*} and c^* independently, only y^{0*} would change as we modify the model. But we still would have the same problem.

An example of this is shown in figure 3.3. The original sample is $\{1, 2^c, 3, 4^c, 5\}$, where ^{c}'s indicate censored observations. Suppose we want to study the survival rate at $t_0 = 3$. There are two different survival estimators in the graph, corresponding to $\hat{F}_\lambda$ for λ equal to 0 and -0.8, solid line and dashed line respectively. The u^*'s are displayed on the left of the vertical axis. Using $\hat{F}_0$, the obtained bootstrap sample is the same as the original one. If we instead use $\hat{F}_{-0.8}$, one observation changes from 3 to 2. But the observation also goes from being a failure to a censored observation. The bootstrap sample is now $\{1, 2^c, 2^c, 4^c, 5\}$. If we diminish λ even more, down to -1.3, the complete sample becomes $\{1, 1, 2^c, 4^c, 5\}$.

In table 3.2 the risk set, number of failures and the KM estimator based on these three bootstrap samples are found. The value of $\hat{\theta}^*$ is highlighted as bold. We see that changing λ from zero to -0.8 causes an increase in $\hat{\theta}^*$. But when changing λ from -0.8 to -1.3, $\hat{\theta}^*$ decreases. These kinds of examples are

		t				
λ		1	2	3	4	5
0	r	5	4	3	2	1
	d	1	0	1	0	1
	$\hat{F}^\star$	0.80	0.80	**0.53**	0.53	0.0
-0.8	r	5	4	2	2	1
	d	1	0	0	0	1
	$\hat{F}^\star$	0.80	0.80	**0.80**	0.80	0.0
-1.3	r	5	3	2	2	1
	d	2	0	0	0	1
	$\hat{F}^\star$	0.60	0.60	**0.60**	0.60	0.0

Table 3.2: Computing $\hat{F}^\star$ using different bootstrap samples.

not hard to find. Especially if there are many censored observations before t_0 in the original data.

This means that we can not perform the bootstrap in the same way as we did with the exact binomial interval. We can not look at one bootstrap sample separately, and calculate a value of λ for which $\hat{\theta}^*$ equals $\hat{\theta}$. We have to look at all bootstrap samples at the same time and find some value of λ for which the distribution of $\hat{\theta}^*$ has shifted its mass enough.

If the number of bootstrap samples is N, we have to save all $N \cdot n$ u^*-variables. For each value of λ of which we want to estimate the probability of $\hat{\theta}^* \leq \hat{\theta}$, we must compute the bootstrap samples as described in figure 3.3. On each bootstrap sample we then calculate $\hat{\theta}^*$ and check how many are above $\hat{\theta}$. This means that we more or less perform a bootstrap for every λ. Thus, this technique involves simple but very extensive calculations.

Actually, since $\hat{\theta}^*$ for some bootstrap samples may not be a monotone function of λ, the number of bootstrap samples for which $\hat{\theta}^*$ is below $\hat{\theta}$ may also be a non-monotone function of λ. Thus, the family of distributions that we obtain by the bootstrap may not form a stochastically ordered family. But the true family of distributions, of which we by the bootstrap have estimated, is a stochastically ordered family. So if the number of bootstrap samples is large enough this will not be a major problem.

3.2 BCa bootstrap interval

In section 2.2.3 the BCa bootstrap interval showed good performance when used on binomial data and using a smoothed bootstrap distribution. As described earlier the BCa method only involves sampling from the cases, with equal sampling weights. But it also involves two adjusting parameters, z_0 and a.

In the same way as done in the binomial case we will use

$$z_0 = \Phi^{-1}\left(\mathsf{P}^*(\hat{\theta}^* < \hat{\theta}) + 0.5\mathsf{P}^*(\hat{\theta}^* = \hat{\theta})\right)$$

instead of the z_0 in (2.8). Thus z_0 is derived using the smoothed distribution rather than the obtained bootstrap distribution as it is.

The acceleration parameter we will use is the non-parametric approximation suggested by Efron (1987) as described in (2.9). The acceleration parameter involves the empirical influence measure of observation $y_1, \ldots, y_n$ on $\hat{\theta}$, $\psi_1, \ldots, \psi_n$ where ψ_i is defined as:

$$\psi_i = \frac{\partial}{\partial \Delta}\left[\hat{\theta}((1-\Delta)F_n + \Delta\xi_i)\right]\Bigg|_{\Delta=0}.$$

In this definition $\hat{\theta}$ is regarded as a function of distributions, and ξ_i is the degenerated distribution with a unit mass at observation y_i. We then get

$$\hat{\theta}((1-\Delta)F_n + \Delta\xi_i) = \prod_j \underbrace{\left(1 - \frac{(1-\Delta)d_j + \Delta \mathrm{I}(y_i = \tau_j, \delta_i = 1)}{(1-\Delta)r_j + \Delta \mathrm{I}(y_i \geq \tau_j)}\right)}_{f_j(\Delta)}$$

where the product is over all j such that $\tau_j \leq t_0$. Then assuming $f_j(\Delta) > 0$ for all j and $\Delta \geq 0$ we get

$$\frac{\partial}{\partial \Delta}\left[\hat{\theta}((1-\Delta)F_n + \Delta\xi_i)\right] = \prod_j f_j(\Delta) \sum_j \frac{f_j'(\Delta)}{f_j(\Delta)}.$$

Evaluating this with $\Delta = 0$ we get the empirical influence measure of observation y_i on $\hat{\theta}$:

$$\psi_i = \prod_j \left(1 - \frac{d_j}{r_j}\right) \sum_j \frac{d_j \mathrm{I}(y_i \geq \tau_j) - r_j \mathrm{I}(y_i = \tau_j, \delta_i = 1)}{r_j(r_j - d_j)}.$$

If $f_j(\Delta) = 0$ for any j we have $\hat{\theta}((1-\Delta)F_n + \Delta\xi_i)$ equal to zero for all $\Delta \geq 0$, and then also $\psi_i = 0$. This will be the case when $\hat{\theta}(F_n) = 0$ and y_i is one of the largest observations, which all are uncensored ones.

However, as with the binomial model of section 2.2, if $\hat{\theta} = 1$, $\hat{\theta}^*$ will have the same value in all bootstrap samples. The same will happen if $\hat{\theta} = 0$ and if there are no censored observations. Thus, we can not use bootstrap in order to compute confidence intervals in these cases. Instead we will use the limits of section 3.1.3, in all cases when $\hat{\theta}$ is either zero or one.

3.3 An illustrative example

To illustrate the types of confidence limits described we use the data of Freireich et al. used by Gehan (1965) and several other authors. Two groups of leukaemia patients of whom the remission time in weeks were recorded. For none of the placebo patients was the remission time censored, making them less interesting for our purpose. But for the group of treated patient some observations were censored. For these 21 patients the recorded remission times were

$$6, 6, 6, 6^c, 7, 9^c, 10, 10^c, 11^c, 13, 16, 17^c, 19^c, 20^c, 22, 23, 25^c, 32^c, 32^c, 24^c, 35^c$$

where the censored observations are indicated by ^{c}. Table 3.3 shows two-sided $1-\alpha$ confidence intervals for the survival rate at two time-points. Bootstrap A is the semiparametric exact interval of section 3.1, Smoothed BCa is the interval described in section 3.2 and the LR interval is the Likelihood ratio interval introduced by Thomas and Grunkemeier (1975). Both bootstrap intervals were computed using 10^5 bootstrap samples.

In all four cases we have the same order between the limits. Looking at the lower limit, the smallest one is always the smoothed BCa limit while LR has the largest limit. The upper limit of smoothed BCa is largest in all four cases and the smallest upper limit is the bootstrap A upper limit. Thus, the smoothed BCa confidence interval covers the other intervals in all four cases. However, the difference between the limits is at most 0.029 and the difference between the LR and bootstrap A limits is never bigger than 0.011.

			Bootstrap A		Smoothed BCa		LR interval	
t_0	$\hat{F}(t_0)$	α	LCL	UCL	LCL	UCL	LCL	UCL
10	0.7529	0.05	0.529	0.899	0.511	0.904	0.540	0.904
		0.10	0.565	0.880	0.554	0.893	0.576	0.885
20	0.6275	0.05	0.389	0.819	0.377	0.835	0.395	0.822
		0.10	0.427	0.793	0.417	0.799	0.432	0.795

Table 3.3: Two-sided 1-α confidence intervals for remission time of leukaemia patients.

3.4 Two simulations

In figure 3.4 and table 3.4 the results of simulations from two different models are shown. The figures show the coverage probability of three different types of intervals: the semiparametric exact interval of section 3.1 (Bootstrap A), the smoothed BCa interval described in section 3.2 (Smoothed BCa) and the Likelihood ratio interval introduced by Thomas and Grunkemeier (1975) (LR).

The simulated samples are of size 50 with the failure times being exponential distributed with mean 1, and the censor variables uniform on $[0, b]$, for b equal to 1.0 and 0.5. The expected number of censored observations in the interval $[0, t]$ for $t \leq b$ is $50 \cdot (1 - F(t))/b$. Since the largest possible value of an observation is b, the graphs show the coverage probability for $F(t)$ between 0 and $F(b) = e^{-b}$.

Actually, the KM is a valid estimator of $F(t)$ only in the interval in which there are observations. Thus the confidence interval is also invalid after the time of the largest observation. The confidence interval used for t above the largest observation in the simulations is $[0, U_n]$, where U_n is the upper limit at the largest observation. In each model and for each confidence interval 10^5 simulations were done.

The Bootstrap A intervals were computed using 10^5 bootstrap samples, and performed as described in section 3.1. Since this method is a very computer intensive one, the cover probability was only computed for 13 and 10 different survival rates in each model respectively, and these corresponds to 23 independent simulations.

The BCa interval was also computed using 10^5 bootstrap samples. The

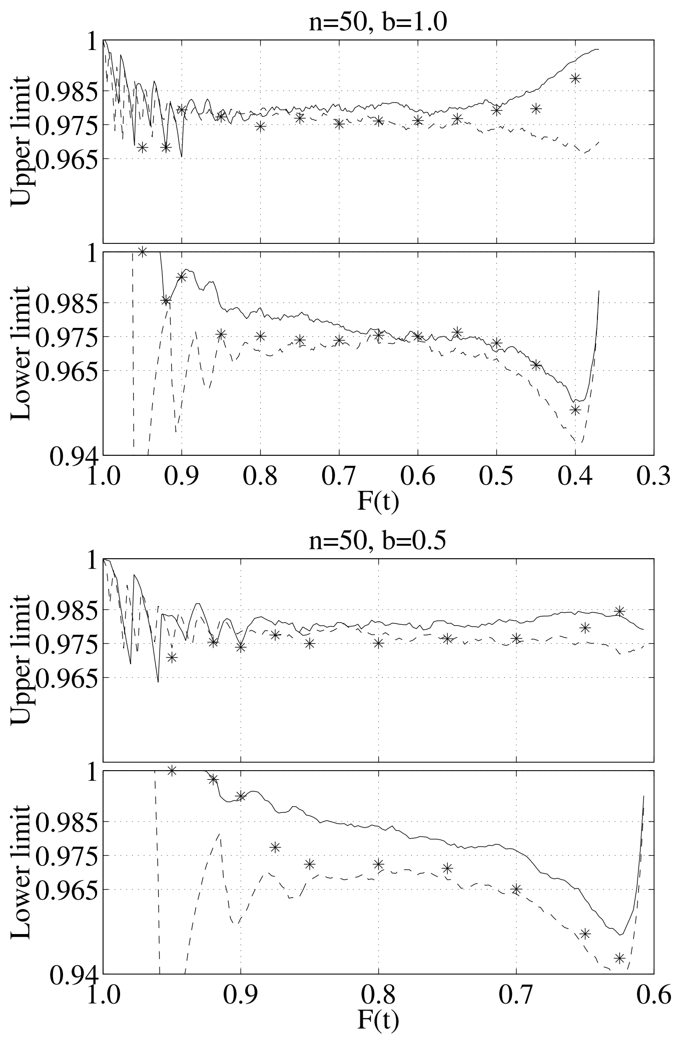

Figure 3.4: Coverage probability of Bootstrap A (*), Smoothed BCa (solid line) and LR (dashed line).

lower and upper limit used when $\hat{\theta}$ equals zero or one, were the ones of bootstrap A. Since it is possible to use the same bootstrap samples for different survival rates with this method, the coverage probability of at least 150 survival rates between 0 and e^{-b} are found in the figures. The coverage probability of the LR interval was computed at the same survival rates as the

		Bootstrap A		Smoothed BCa		LR interval	
b	$F(t)$	α_l	α_u	α_l	α_u	α_l	α_u
1.0	0.950	*0.00*	3.17	*0.00*	1.49	8.65	**2.40**
	0.920	1.43	3.17	1.56	3.29	*1.89*	*2.05*
	0.900	*0.75*	*2.06*	0.67	3.45	4.47	2.04
	0.850	**2.43**	**2.27**	1.63	2.10	**2.43**	**2.21**
	0.800	**2.49**	**2.55**	1.65	2.16	2.97	2.08
	0.750	**2.60**	**2.31**	1.83	2.02	**2.81**	2.19
	0.700	**2.61**	**2.48**	2.15	1.92	**2.74**	**2.44**
	0.650	**2.46**	**2.39**	**2.27**	1.92	**2.70**	**2.31**
	0.600	**2.49**	**2.38**	**2.54**	2.11	**2.62**	**2.48**
	0.550	**2.37**	**2.32**	**2.45**	2.05	2.87	**2.44**
	0.500	**2.69**	2.08	2.96	1.91	3.15	**2.70**
	0.450	*3.34*	2.03	*3.34*	1.46	3.96	*2.86*
	0.400	4.66	1.14	*4.35*	0.62	5.59	*3.15*
0.5	0.950	*0.00*	2.91	*0.00*	1.67	9.21	**2.63**
	0.920	0.26	**2.46**	0.19	**2.41**	**2.26**	**2.26**
	0.900	0.75	**2.61**	*0.85*	**2.55**	4.35	2.15
	0.875	**2.26**	**2.25**	1.18	1.96	3.12	**2.28**
	0.800	**2.76**	**2.49**	1.64	1.85	3.02	**2.24**
	0.850	**2.76**	**2.50**	1.33	2.13	3.23	**2.21**
	0.750	2.88	**2.36**	*2.16*	1.91	3.22	**2.29**
	0.700	3.49	**2.35**	**2.36**	1.83	3.63	**2.52**
	0.650	4.81	2.04	*3.80*	1.60	5.23	**2.44**
	0.625	5.53	1.55	*4.85*	1.67	6.06	**2.81**

Table 3.4: $\mathsf{P}(\text{lower limit} > F(t)) \cdot 100\% = \alpha_l$ and $\mathsf{P}(\text{upper limit} < F(t)) \cdot 100\% = \alpha_u$ of Bootstrap A, Smoothed BCa and LR.

BCa. Thus, the whole coverage probability curve in each graph corresponds to the same simulations.

All intervals are symmetric and at level 0.95, i.e. the coverage probability of the lower and upper limits should ideally be 0.975.

In all simulations the lower limit at the last failure is carried forward up to the last censored observation and beyond this point set to zero. The coverage probability of all three lower limits decreases at the lower range of

$F(t)$. This might be a consequence of the way we have defined the lower limit. If we instead had chosen to set the limit to zero already after the last failure the picture might have been different.

Comparing smoothed BCa and LR the former is more conservative overall. Only for $F(t)$ above 0.9 is the coverage probability of the upper BCa limit less than the one of LR. But the differences for the upper limit between these two is very small except for $F(t) < 0.5$ and $b = 1.0$. The difference in coverage probability of the lower limits is more pronounced. The coverage of the lower LR limit is almost always below the intended limit. The coverage probability of the lower BCa limit is too high for large survival rates, and drops below the intended level for larger values, with a minimum of 0.96 and 0.95 for b equal to 1.0 and 0.5 respectively. The minimum of the LR limit in the same region of $F(t)$ is 0.945 and 0.94 (outside graph). The lowest coverage probability of the lower LR limit is for $F(t)$ around 0.95. Around this point is the coverage less than 0.9 for both values of b.

The confidence limits of bootstrap A were simulated for 23 cases. For each of these cases the corresponding values of the other intervals together with the ones of bootstrap A are found in table 3.4. Instead of coverage probabilities the estimated probabilities of a limit failing to cover the parameter are tabulated. A probability that lies between 2.2% and 2.83% is highlighted as bold. The chosen range is the exact binomial interval of $\hat{p} = 0.025$ when the number of trials is 10 000. Across each row, separately for lower and upper limit, if there is no coverage probability within the range, the probability closest to 2.5% is in *italic*.

As seen in the table and in the figure for $F(t) \geq 0.9$ the coverage of the lower bootstrap A limit is very close to BCa. This is not surprising since for $\hat{F}(t_0)$ equals one, they both use the same limit. In this region is the lower bootstrap A limit rather conservative, but below 0.9 and down to the point of 50% expected censoring ($F(t)$ equal to 0.5 and 0.75 respectively) bootstrap A lower limit has probability within the range $[2.2\%, 2.83\%]$ in 11 of 12 cases.

Looking at the bootstrap A upper limit the difference compared to LR is not very big, the exception being for $F(t) < 0.5$ and $b = 1.0$ where the coverage probability goes up in the same way as smoothed BCa does.

The Smoothed BCa is overall quite conservative, except for small $F(t)$ but it is questionable if we would see the same drop in coverage probability with the lower limit equal to zero beyond the last failure. With the current definition of the lower limit it is only in this region that the lower Smoothed BCa limit looks like a winner, since the other has lower coverage probability.

The best lower limit judging from this simulations is the one of Bootstrap A. But for the upper limit it is harder to judge in favour of Bootstrap A or LR, the upper LR limit seems better for extreme survival rates while Bootstrap A seems better for more moderate ones.

3.5 Summary

In section 3.1 we introduced the semiparametric exact confidence interval bootstrap A which we computed using bootstrap. It is based on the constrained KM estimator used in the LR interval of Thomas and Grunkemeier (1975). We have also shown how the BCa interval can be calculated using a smoothed bootstrap distribution in the same way as in section 2.2.3.

These two intervals, bootstrap A and smoothed BCa, together with the LR interval was compared in two simulation in section 3.4. Judging from these simulations the smoothed BCa is more conservative than the other two, especially the lower limit. The lower LR limit showed too low coverage probability, while the upper LR and bootstrap A limit seems to be equally good. Since the lower bootstrap A limit was the lower limit with coverage probability closest to the specified level, and the upper limit was as good as LR, bootstrap A seems to be better than the other two in general.

Chapter 4

Group-wise dependent survival data

In this chapter we will consider the random censoring model where groups of observations no longer are mutually independent. We assume that there are m groups each sized n. All mn failure times still have the same marginal distribution, and the failure times are independent between groups. But within group we assume that there is a positive dependence, and the dependence is the same for all pairs within a group, across all groups. We do not want to make any further assumption of dependency within group, only that it may exists and that it is positive. Further, we assume that the censoring is univariate, i.e. all censored observations of a group are censored at the same time.

We start this chapter by discussing what we ideally would require of the sampling model. With this in mind we take a look at the analogue of bootstrap method 1 and 2 in section 2.4. We then briefly discuss frailty models and how these can be used in our setting.

4.1 Preferred sampling model

Formally, for $i = 1, \ldots, m$ $(Y_{i1}^0, \ldots, Y_{in}^0)$ is distributed according to the multivariate survival function F_0 and C_i is distributed according to the univariate survival function G. All marginals of F_0 are equal to F, e.g. $F_0(0, \ldots, t) = F(t)$, and the coordinates are exchangeable, e.g. $F_0(t_1, t_2, \ldots, t_n) = F_0(t_n, \ldots, t_2, t_1)$. We observe $Y_{ij} = \min(Y_{ij}^0, C_i)$ and

$\delta_{ij} = I(Y_{ij}^0 \leq C_i)$, let (y_{ij}, δ_{ij}) be the observed value.

Ying and Wei (1994) studied the KM estimator under models with dependent observations. They concluded that KM is still consistent under rather mild assumptions. For the model of this chapter these assumptions are fulfilled. Thus KM is still a valid estimator of the survival function F.

There are many KM-type estimators of G based on the kind of data considered here. Tsai and Crowley (1998) proposed an estimator of the bivariate survival function but they also studied four different KM estimators of G. They did not assume the same marginal for the two coordinates, but their censoring was univariate as in our model. The four estimators of G were based on either maximum of y_{i1}, y_{i2}, minimum of y_{i1}, y_{i2}, or only one of the coordinates.

They concluded that all four estimators are consistent estimators of G, but the one with the smallest asymptotic variance is the one based on maximum of y_{i1}, y_{i2}. The one with the biggest asymptotic variance is the one based on minimum of y_{i1}, y_{i2}, and the other two lie in between.

This result is not surprising looking at the case when the smallest of y_{i1} and y_{i2} is uncensored and the other is censored. For simplicity assume $y_{i1} < y_{i2}$. Then, when using maximum of y_{i1}, y_{i2}, the information obtained about the censor variable C_i, is that is equal to y_{i2}, i.e. for estimating G we have an uncensored observation equal to y_{i2}. If we instead use minimum of y_{i1}, y_{i2}, we only know that C_i is greater or equal to y_{i1}, that is, for estimating G we have a censored observation equal to y_{i1}.

Hence, to proceed in the same manner as in the independent case we would prefer sampling techniques where the survival times (Y_{ij}^{*0}) have marginal distribution according to the KM estimator of F. Regarding the dependency the sampling technique should take it into account in an unbiased way, but otherwise we do not want to impose any structure on the dependency. We also would prefer techniques where the survival function of the censor variable (C_i^*) is the KM estimator of G based on maximum of y_{ij} over the coordinates.

4.2 A solution

Let there be k distinct times of failures and censored observations, $\tau_1, \ldots, \tau_k$, among the mn observations. If there are ties between failures and censored observations, the censored observations are taken to occur immediately after the failures, and correspond to different τ's. For $i = 1, \ldots, m$, let $r_{is} = \sum_j I(y_{ij} \geq \tau_s)$ be the number of observations at risk at τ_s of group i, d_{is} and c_{is} is the number of failures and censored observations at τ_s respectively. Note that $d_{is} \cdot c_{is} = 0$ and $d_{\cdot s} \cdot c_{\cdot s} = 0$ for all i and s, where e.g. $d_{\cdot s}$ is the sum of d_{is} for $i = 1, \ldots, m$. The estimator of the marginal survival function F and the preferred estimator of G can be written as

$$\hat{F}(t) = \prod_{s:\tau_s \leq t} \left(1 - \frac{\sum_i d_{is}}{\sum_i r_{is}}\right) \tag{4.1}$$

$$\hat{G}(t) = \prod_{s:\tau_s \leq t} \left(1 - \frac{\sum_i \mathrm{I}(c_{is} > 0)}{\sum_i \mathrm{I}(r_{is} > 0)}\right) \tag{4.2}$$

since we get the total number of observations at risk and total numbers of failures and censored observations by summing over the groups. For time being, assume that $\hat{F}(\tau_s) = 0$, i.e. the largest observation is a failure.

For $s = 1, \ldots, k$ and $i = 1, \ldots, m$ define

$$\begin{aligned} f(\tau_s) &= \hat{F}(\tau_s-) - \hat{F}(\tau_s) \\ \nu(i) &= \textstyle\sum_{s:d_{\cdot s}>0} f(\tau_s)\frac{d_{is}}{d_{\cdot s}} \\ f_i(\tau_s) &= \begin{cases} f(\tau_s)\frac{d_{is}}{d_{\cdot s}\nu(i)} & \text{if } d_{\cdot s} > 0 \text{ and } \nu(i) > 0 \\ 0 & \text{otherwise} \end{cases} . \end{aligned} \tag{4.3}$$

Hence, f is the frequency function of $\hat{F}$ on $\tau_1, \ldots, \tau_k$. For each i with $\nu(i) > 0$ f_i is also a frequency function on $\tau_1, \ldots, \tau_k$, and ν is a frequency function on the indexes 1 to m.

Now consider the following way of sampling $Y_{11}^{*0}, \ldots, Y_{1n}^{*0}$: First draw a group according to the frequency function ν. If i^* is the selected group, $y_{11}^{*0}, \ldots, y_{1n}^{*0}$ is obtained by sampling n times according to the frequency function f_{i^*} on $\tau_1, \ldots, \tau_k$. These two steps are repeated m times to get a bootstrap sample from m groups. Thus first we do a weighted sampling from the m groups. For each time we select group i we then perform a weighted sampling from the failures of group i. The sampling of the m censor variables is performed as before using $\hat{G}$. It is easy to check that this way of sampling has failure times with marginal distribution corresponding to $\hat{F}$.

If $\hat{F}(\tau_s) > 0$, i.e. the largest observation is a censored one, neither f nor ν as defined above are frequency functions. But if we add one failure time, τ_{k+1}, greater than all original ones to the sample space of f with $f(\tau_{k+1}) = \hat{F}(\tau_s)$, f is a true frequency function . We do the same thing with ν; we add an index, $m+1$, with $1-\nu(m+1)$ equal to the sum of $\nu(i)$ for $i = 1, \ldots, m$. Finally by setting $f_{m+1}(k+1)$ equal to 1, (f_{m+1} zero otherwise) we have the correct marginal distribution of the sampled observations.

The nice part with this technique is that we may use another marginal function when defining the frequency functions of (4.3), e.g. the constrained KM estimator F_λ used in the independent case. Hence we could calculate semiparametric confidence intervals as done in section 3.1. However, as we will see below this is shadowed by two major drawbacks.

4.2.1 Drawback no 1

Suppose there are no censored observations in the original sample. Then all bootstrap samples will also contain failures only. Thus, both $\hat{F}(t_0)$ and $\hat{F}^*(t_0)$ are just the relative frequencies of observations greater than t_0 in the original and bootstrap sample respectively. Further, since $\nu(i) = 1/m$ the groups are sampled with equal weights in step 1.

In the second step we sample from the observations of the group selected in step 1. Since $f_i(s) = d_{is}/n = \#\{j : y_{ij} = \tau_s\}/n$ for all i and s we do this with equal weights. This means that the number of observations greater than t_0 is binomial(n, p) distributed with p equal to the relative frequency of observations greater than t_0 of the selected group.

This is exactly the situation studied in section 2.4. There we found that the bootstrap distribution will on average be too wide. When there actually is no dependency, the variance can be twice as big as it should be.

Consider the artificial data in the left graph of figure 4.1. The data consists of 9 bivariate observations where the failure times are either short medium or long. The observations are uniformly spread over the 9 squares. Hence, the data suggest no dependence. The bivariate distribution obtained by using this bootstrap technique is shown in the right graph. The bootstrap distribution shows a relatively large dependency with Kendall's coefficient of concordance equals 0.47. Generalising this example to k categories and observations uniformly spread out over the k^2 bivariate cases Kendall's coefficient of concordance equals $5/12 + 1/6k$, thus the concordance never falls below 0.415.

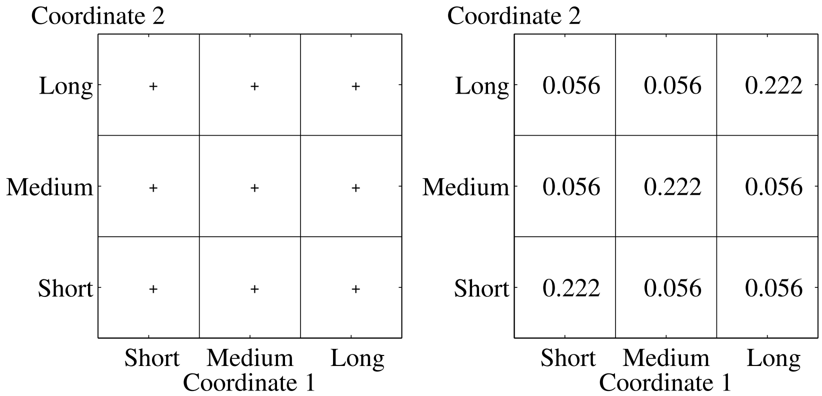

Figure 4.1: Artificial bivariate data (left graph) and the bivariate bootstrap distribution of section 4.2 (right graph).

4.2.2 Drawback no 2

But the censoring itself also contributes to the dependency being too big in the bootstrap distribution. If for instance all but one observation of a group is censored the conditional distribution f_i is degenerated at the failure. We illustrate this by another example. In this example the failure times are independent and uniformly distributed on $\{1, 3, 5\}$, thus the bivariate distribution is uniformly on $\{1, 3, 5\}^2$. The distribution of the univariate censor variables is uniform on $\{2, 4, 6\}$. Suppose we have a "typical" sample from this model where the proportions of each possible case is the same as we would have in an infinite sample.

Then the proportion of (Y_1, Y_2) equals (2,2) in the typical sample will be $1/3 \cdot 1/3 \cdot 1$, and the proportion equals (6,6) will be $1/3 \cdot 1/3 \cdot 1/3$. The proportions of each of the 11 different values of (Y_1, Y_2), disregarding the order between the coordinates, are found in the upper part of table 4.1. The censored values are marked with a ^{c}.

Note that over the two coordinates the proportions of failures and censorings at t equals $\{2, 3, 4, 5, 6\}$ is proportional to $\{18, 0, 12, 0, 6\}$ and $\{0, 12, 0, 6, 0\}$ respectively. Thus the KM estimator of F based on the sample is the uniform distribution on $\{1, 3, 5\}$. Further, when looking at maximum over the coordinates the corresponding numbers are $\{3, 0, 6, 0, 5\}$ and $\{0, 8, 0, 5, 0\}$ respectively. Thus the KM estimator of G is the uniform distri-

Y_1	2	2	2	2	2	3^c	4	4	4	5^c	6
Y_2	2	3^c	4	5^c	6	3^c	4	5^c	6	5^c	6
Proportion·27	3	4	4	2	2	4	2	2	2	1	1

i	1	2	3	4	5	6	7	8	9	10	11
$\nu(i) \cdot 54$	6	4	10	2	8	0	6	3	9	0	6

s	$f_i(s) \cdot 60$										
2	60	60	24	60	15	0	0	0	0	0	0
4	0	0	36	0	0	0	60	60	20	0	0
6	0	0	0	0	45	0	0	0	40	0	60

S^*	2	4	6
6	0.028	0.037	0.269
4	0.044	0.252	0.037
2	0.261	0.044	0.028

Table 4.1: Proportion of the 11 possible cases in the typical sample, the frequency functions of the two-stage sampling technique and the bivariate bootstrap distribution function S^*.

bution on $\{2, 4, \infty\}$.

In the middle part of table 4.1 the frequency functions ν and $f_1, \ldots, f_{11}$ are shown. As defined there is one conditional distribution for each group, but if groups have the same set of observations we might as well consider them as one group when we sample. Thus in this example there will be 11 distinct groups.

The last part of the table display the bivariate distribution of the bootstrap. The bootstrap distribution of figure 4.1 is the distribution we get if the probability of a censored observation is equal to zero. Comparing these two bootstrap distributions the one in table 4.1 is even more concentrated on the diagonal than the one of figure 4.1. Kendall's coefficient of concordance equals 0.68 compared to 0.47 in the case of no censoring.

Now we generalise by replacing the distributions of the typical sample with uniform on $\{2, 4, \ldots, 2k\}$ and $\{1, 3, \ldots, 2k-1\}^2$ for the censor variable and the failure times respectively. Then for k equals 5, 10 and 50 Kendall's coefficient of concordance of the bootstrap distribution equals 0.69, 0.71 and

0.72 respectively. The probability of an observation being censored is $(k-1)/2k$, i.e. for $k = 3$ this equals $1/3$ but for $k = 10$ the probability is 0.45. Hence as we increase k the proportion of censored observations increase which explain the increased Kendall's coefficient. If we instead let the censor distribution be such that the probability for a censored observation is $1/3$ for all k, by setting $\mathsf{P}(C = 2k) = 1/3$ and $\mathsf{P}(C = j) = 2/(3(k-1))$ for $j < 2k$, Kendall's coefficient equals 0.5923, 0.5739 and 0.5630. Hence instead we see a slow decline but still at k equals 50 the number is fairly high. If the probability for a censored observation is equal to zero the corresponding numbers are 0.45, 0.43 and 0.42. This gives an indication that the enlarged dependency in the bootstrap caused by the censoring will be fairly large also for continuous data.

4.3 Sampling groups

There are at least two ways of sampling groups. The most straightforward way is to sample the observations from the groups as they are, that is, to sample Y^* directly from the cases we have observed. Another way would be to sample failure times and censorings separately. But then one has to decide how to treat the case when a group has for instance only one failure. We will only consider the former way of sampling.

Suppose n equals 2, i.e. we have observations from a bivariate survival model. When we sample (Y_1^*, Y_2^*) from the m groups we have

$$S_0(t,s) \;=\; \mathsf{P}^*\left(Y_1^* > t, Y_2^* > s\right) \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}(y_{i1} > t, y_{i2} > s).$$

The criteria of our preferred sampling technique would be fulfilled if the product $\hat{F}(t)\cdot\hat{G}(t)$ was equal to $(S_0(t,0)+S_0(0,t))/2$, where $\hat{F}$ and $\hat{G}$ are the ones in (4.1) and (4.2) respectively. But this does not hold. Instead, for s such that $F(\tau_s) > 0$ we have

$$\frac{S_0(\tau_s,0)+S_0(0,\tau_s)}{2\hat{F}(\tau_s)} \;=\; \frac{\sum_{i=1}^m \mathrm{I}(y_{i1} > \tau_s) + \mathrm{I}(y_{i2} > \tau_s)}{2m\hat{F}(\tau_s)} \;=\; \frac{\sum_{i=1}^m r_{i,s+1}}{2m\hat{F}(\tau_s)}.$$

The sum equals $r_{\cdot s+1}$ and using $r_{\cdot 1} = 2m$ and the fact that $d_{\cdot j}c_{\cdot j} = 0$ we get

$$\frac{r_{\cdot s+1}}{r_{\cdot 1}\hat{F}(\tau_s)} \;=\; \frac{r_{\cdot s+1}}{r_{\cdot 1}}\prod_{j=1}^{s}\left(1-\frac{c_{\cdot j}}{r_{\cdot j}-d_{\cdot j}}\right) \;=\; \prod_{j=1}^{s}\left(1-\frac{c_{\cdot j}}{r_{\cdot j}}\right). \tag{4.4}$$

Since this equals the KM estimator of G treating the univariate censoring as bivariate and independent we are outside the criteria of our "preferred model". Not only is the distribution wrong, this estimator of G is also not sufficient when the failure times within groups are dependent.

Can we perform a weighted sampling from the groups that will correspond to the preferred model, i.e. can we find weights $\nu_1, \ldots, \nu_m$ such that

$$\frac{1}{2}\sum_{i=1}^{m} \nu_i r_{i,s+1} = \hat{F}(\tau_s)\hat{G}(\tau_s)$$

for $s = 1, \ldots, k-1$? If the failure times and censor variables are genuine continuous variables, the only tied observations will be within a group and only between censored ones. Thus, as soon as two groups has only one failure each we have $k - 1 > m$. Then this equation system has more equations than unknowns and examples with no solution are not hard to find. Hence, in general it will not be possible to do a weighted sampling from the groups that will correspond to the preferred model. The fact that there are cases with no solution to the equation system means that it neither will be possible to compute the type of semiparametric confidence interval of section 3.1.

4.4 Frailty models

Judging from the previous sections it seems as if without tightening the preferred model, e.g. the structure of the dependency, or altering the model in some way it is hard to find a good sampling model.

An appealing way of structuring the dependency is to consider frailty models introduced by Vaupel, Manton, and Stallard (1979). In this section we will briefly go into the frailty model and discuss how one may use this for our purpose. For a more thorough description of frailty models and various characteristics of them see e.g. Hougaard (2000).

4.4.1 Continuous failure distribution

The frailty model is a mixture model where the mixture term is the frailty common to the failure times of a group and constant over time. The frailty is a non-negative random variable distributed according to H. Conditional on the frailty,η_i, the failure times of group i are independent and the hazard function for $Y_{i1}^0, \ldots, Y_{in}^0$ is assumed to be of the form

$$\eta_i \omega(t). \tag{4.5}$$

The function $\omega(t)$ is common to all groups and can be interpreted as a baseline hazard. Hence, the conditional survival function is

$$F_{|\eta_i}(t) = \exp\{-\eta_i \int_0^t \omega(u)du\} = \exp\{-\eta_i \Omega(t)\}.$$

The unconditional survival function is then obtained by integrating η_i out:

$$F(t) = \int_0^\infty \exp\{-x\Omega(t)\}dH(x) = L_H(\Omega(t)),$$

i.e. F is the Laplace transformation of the frailty distribution L_H, evaluated at the integrated baseline hazard function $\Omega(t)$. The bivariate survival function is also computed using L_H:

$$F_0(t_1, \ldots, t_n) = L_H(\sum_{j=1}^n \Omega(t_j)).$$

This way of describing the model is known as the conditional parameterization, but since our goal is to estimate the marginal survival function F it is more convenient to use the marginal parameterization

$$F_0(t_1, \ldots, t_n) = L_H\left[\sum_{j=1}^n L_H^{-1}(F(t_j))\right].$$

The frailty model is commonly used with the frailty distribution being specified, e.g. gamma or positive stable, and the baseline hazard unspecified. With the frailty distribution gamma(γ,γ) the likelihood contribution of group i is

$$\left[\prod_{j=1}^n \lambda(y_{ij})^{\delta_{ij}} \exp\{\delta_{ij}\Lambda(y_{ij})/\gamma\}\right] \cdot \left[\sum_{j=1}^n \exp\{\Lambda(y_{ij})/\gamma\} - (n-1)\right]^{-(\gamma+d_{i\cdot})} \frac{\gamma^{-d_{i\cdot}}\Gamma(\gamma+d_{i\cdot})}{\Gamma(\gamma)} \tag{4.6}$$

where λ and Λ are the hazard and integrated hazard of F respectively. Since λ is unspecified the likelihood is maximized by using one parameter for each distinct failure time. That is, for each τ_s corresponding to a failure time $\lambda(\tau_s) = \lambda_s$ and $\Lambda(\tau_s)$ is the sum of $\lambda_1, \ldots, \lambda_s$. If no failures occur at $t = \tau_s$ $\lambda(\tau_s) = 0$.

When fitting this kind of model the estimator of F is of course no longer the KM of (4.1). The function being estimated is the integrated hazard and then we use $\exp\{-\hat{\Lambda}(t)\}$ as the estimator of F. Hence the estimator of F will be closer to the estimator based on the Nelson-Aalen estimate of Λ:

$$\hat{F}_{NA}(t) = \exp\{-\sum_{s:\tau_s \le t} \frac{\sum_i d_{is}}{\sum_i r_{is}}\} = \exp\{-\sum_{s:\tau_s \le t} \frac{d_{\cdot s}}{r_{\cdot s}}\}. \tag{4.7}$$

Comparing (4.7) and (4.1) the difference will not be large if $d_{\cdot s}/r_{\cdot s}$ is sufficiently small. Thus, for large survival rates the difference will not be large but as the rates gets smaller the difference will increase. Actually $\hat{F}_{NA}$ is always strictly positive, even if the last observation is a failure, while $\hat{F}$ equals zero at the time of the last observation when this is a failure.

By inserting the KM estimator $\hat{F}$ into the likelihood and maximize it with respect to the parameters of the frailty distribution only, the criterion of the preferred model are fulfilled. The problem with $\hat{F}$ equals zero corresponding to Λ equals ∞ could be solved by holding $F(t)$ equal to $\hat{F}(t)$ only for $t \le t_0$, or by treating all failures after t_0 as censored just after t_0. Either way we instead will have different estimators of the frailty parameters for different t_0's. The case when $\hat{F}(t_0)$ equals zero still remains to be handled.

4.4.2 Discrete failure distribution

As described above the frailty model is best suited for absolutely continuous distributions. If the data is discrete, or estimators closer to the KM are more pleasing, the frailty can instead of (4.5) enter the model in the same way as the regression function in the discrete proportional hazard model. That is,

$$\frac{h_{s|\eta_i}}{1 - h_{s|\eta_i}} = \eta_i \frac{h_s}{1 - h_s} \tag{4.8}$$

where $h_{s|\eta_i}$ and h_s are the conditional and baseline discrete hazard functions at τ_s respectively. With the frailty distribution H the unconditional survival function is

$$F(t) \quad = \quad \int_0^\infty \left[\prod_{s:\tau_s \le t} (1 - h_{s|x}) \right] dH(x) \tag{4.9}$$

$$= \int_0^\infty \left[\prod_{s:\tau_s \le t} \frac{1-h_s}{1-h_s+xh_s} \right] dH(x) \tag{4.10}$$

and the contribution of group i to the likelihood is

$$\left[\prod_{s=1}^{k} \binom{r_{is}}{d_{is}} h_s^{d_{is}} (1-h_s)^{r_{is}-d_{is}} \right] \cdot \int_0^\infty \left[\prod_{s=1}^{k} \frac{x^{d_{is}}}{(1-h_s+xh_s)^{r_{is}}} \right] dH(x). \tag{4.11}$$

This likelihood is harder to evaluate and maximize than (4.6). Moreover, a marginal parameterization of the model would be even worse. Hence to have a model that is within the preferred model, i.e. requiring (4.9) equals $\hat{F}(t)$ for every t, makes the evaluation very hard.

With the distribution for the frailty set to gamma(γ,γ), the discrete frailty model above was fitted to the artificial data of figure 4.1 and to the typical data set of table 4.1. The likelihood functions are maximized when γ is set to ∞ in both cases. Hence in the fitted models all observations are independent. The marginals of the fitted model equal the uniform distribution on {short,medium,long} and $\{1, 3, 5\}$ respectively. Hence in both cases we end up with the preferred marginal distribution and the coordinates are independent, as they should.

4.4.3 Summary

Perhaps the most obvious way of sampling in two steps, first group and then within group, is to sample with equal weights in both steps and straight from the cases. This way of sampling leads to the same result as when sampling groups only. Namely that the distribution of the censor variables equals (4.4), which is not a sufficient estimator. But what probably is more important, the dependency of the sampling distribution would be too high as with the two-step sampling technique described in section 4.2. This downside makes these two solutions more or less useless since they are very conservative unless the dependency is rather strong.

In case of no censored observations sampling only groups will correspond to the preferred model. But in presence of censoring the sampling distribution of the censor variable is not a sufficient estimator of G. This drawback may not be that important, but still we can not calculate a semiparametric interval with this way of sampling. However, calculating a BCa interval, or other bootstrap intervals based on sampling cases, treating the groups as the experimental unit is quite straightforward and as easy as with independent observations.

The usage of continuous frailty models inevitably leads to the Nelson-Aalen based estimator of F. With this estimator replacing KM as the estimator of the marginal failure distribution F the continuous frailty model is probably the way to go. Instead of using the constrained version of KM in (3.1) when calculating a semiparametric confidence interval a constrained estimator of the integrated hazard function could then be used. Another solution would be to use a proportional hazard model.

The discrete frailty model is an interesting solution when there are many ties in the data. We have not found any references about this in the literature. Furthermore, since the fitting of such a model, especially when holding the marginal fix, is indeed a challenge for the programmer, it must be considered to be an outsider.

Chapter 5

Concluding remarks

An exact semiparametric confidence interval (bootstrap A) for the survival rate calculated using bootstrap has been introduced. It is based on the constrained KM estimator, and assumes that all observations are independent. Besides the bootstrap approximation it coincides with the exact binomial interval when there are no censored observations. In presence of censored observations it is exact against alternatives given by constrained maximum likelihood estimators of the survival function.

The accelerated bias corrected percentile interval (BCa) when used on binomial data was shown to have coverage probabilities not far from the exact interval. It was more conservative for p close to zero or one compared to the exact interval, while the coverage probability was slightly too low for moderate values of p. A smoothed bootstrap distribution was used.

The performance of the bootstrap A and smoothed BCa intervals together with the likelihood ratio interval of Thomas and Grunkemeier (1975) (LR) in case of censored observations was compared in two simulations. For the upper limits the differences in coverage probability were small. However, smoothed BCa was slightly more conservative than the other two.

Between the lower limits there were greater differences. Smoothed BCa was in general more conservative than the other two and LR had too low coverage probability. The coverage probability of the lower bootstrap A was in between smoothed BCa and LR and except for survival rates above 0.9 it was the limit closest to the intended level.

Besides the choice of alternatives, the way of sampling from different models in order to compute the semiparametric confidence interval was quite straightforward in the case of independent observations. When groups of observations are assumed positively correlated it is difficult even to find a way of sampling from the null hypothesis, i.e. from an unconstrained model suggested by the data. Actually, at present there is no final solution in the literature to the problem of finding a non-parametric estimator of the bivariate survival function. Hence, without making any stronger assumptions we have not found any satisfying sampling model.

One way of solving a specified problem is always to change the specification. In our case this could be done by assuming a model for the dependency structure instead of letting the structure be determined completely by the data. A well-documented way of doing this is the concept of frailty. It would then be more natural to base the inference of the survival function on the Nelson-Aalen based estimator instead of the KM estimator.

References

Akritas, M. G. (1986). Bootstrapping the Kaplan-Meier estimator. *J. Amer. Statist. Assoc.*, *81*(396), 1032–1038.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, *74*(3), 457–468.

Böhmer, P. E. (1912). Theorie der unabhängigen wahrscheinlichkeiten. In *Rapports mémories et procès-verbaux de septième congrès international d' actuaries, Amsterdam* (Vol. 2, pp. 327–343).

Collett, D. (1991). *Modelling binary data.* London: Chapman & Hall.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge: Cambridge University Press. (With 1 IBM-PC floppy disk (3.5 inch; HD))

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, *7*(1), 1–26.

Efron, B. (1981a). Censored data and the bootstrap. *J. Amer. Statist. Assoc.*, *76*(374), 312–319.

Efron, B. (1981b). Nonparametric standard errors and confidence intervals. *Canad. J. Statist.*, *9*(2), 139–172. (With discussion and a reply by the author)

Efron, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.*, *82*(397), 171–200. (With comments and a rejoinder by the author)

Eriksson, B., & Adell, R. (1994). On the analysis of life tables for dependent observations. *Statistics in Medicine*, *13*(1), 43–51.

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

Hjort, N. L. (1985). *Bootstrapping Cox's regression model* (Tech. Rep. No. NSF-241). Stanford, California: Department of Statistics.

Hougaard, P. (2000). *Analysis of multivariate survival data.* New York: Springer-Verlag.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, *53*, 457–481.

Kim, J.-H. (1995). Conditional bootstrap methods for censored survival data. *J. Korean Statist. Soc.*, *24*(1), 197–218.

Reid, N. (1981). Estimating the median survival time. *Biometrika*, *68*(3), 601–608.

Thomas, D. R., & Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.*, *70*(352), 865–871.

Tsai, W.-Y., & Crowley, J. (1998). A note on nonparametric estimators of the bivariate survival function under univariate censoring. *Biometrika*, *85*(3), 573–580.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual fraility on the dynamics of mortiality. *Demography*, *16*, 439–454.

Ying, Z., & Wei, L. J. (1994). The Kaplan-Meier estimate for dependent failure time observations. *J. Multivariate Anal.*, *50*(1), 17–29.