A review of inferential methods for the Kaplan-Meier estimator

Stuart Barber and Christopher Jennison

SUMMARY

We survey existing methods of inference using the Kaplan-Meier estimator, focusing on testing $H_0: S(t^*) = p^*$ for some chosen $t^*$ and $p^*$. Performance of these methods is assessed by simulation. The problems these methods encounter are analysed and the reasons for these problems discussed.

1 Introduction

1.1 The problem

A fundamental problem in survival analysis is estimating survival probabilities, as either a point estimate or a confidence interval. If a parametric model is assumed, such as assuming exponential or lognormal distributions for failure times, then estimation of the model parameters can be done using maximum-likelihood methods, modified to deal with the problem of censored data.

Kaplan and Meier (1958) proposed a nonparametric method which has been accepted as the standard estimator for such probabilities. Since this estimate was proposed, there have been many methods suggested for carrying out inference using the Kaplan-Meier estimator. In this report we survey these methods and assess their performance by simulation.

1.2 Notation and definitions

Consider i.i.d. failure times $x_1, \ldots, x_n$ subject to right-censoring times $y_1, \ldots, y_n$. Then the observed data consist of survival times $\{(t_i, \delta_i) : i = 1, \ldots, n\}$ where $t_i = \min(x_i, y_i)$ and $\delta_i = I(t_i = x_i)$. Let $r$ be the number of distinct failure times, which are ordered as $t_{(1)}, \ldots, t_{(r)}$. Define $n_j$ to be the number of individuals surviving up to time $t_{(j)}$, i.e. the number of individuals at risk at time $t_{(j)}$, and $d_j$ as the number of failures at time $t_{(j)}$. 

To estimate the survivor function \( S(t) = \mathbb{P}\{X_i > t\} \), we use the Kaplan-Meier estimator (Kaplan and Meier, 1958).

\[
\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \frac{n_j - d_j}{n_j}, \quad 0 \leq t \leq \max_i \{t_i\}.
\]

It can be shown that the Kaplan-Meier estimator is asymptotically normal with mean \( S(t) \) and variance to be estimated (Meier, 1975). Variance estimation can be done by several methods. The first is Greenwood’s formula (Greenwood 1926);

\[
\hat{\sigma}_G^2(t) = \{\hat{S}(t)\}^2 \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}.
\]

This is derived by estimating each term in the product expansion of \( \hat{S}(t) \) separately. An alternative is from Peto at. al. (1977)

\[
\hat{\sigma}_P^2(t) = \{\hat{S}(t)\}^2 \{1 - \hat{S}(t)\} / n_k, \quad t_{(k)} \leq t < t_{(k+1)},
\]
suggested for use when \( \hat{S}(t) \) is high or low. Simon and Lee (1982) suggest replacing \( n_k \) by \( n_{k+1} \) in this expression; with this change the variance estimation reduces to the binomial variance in the absence of censoring. However, this estimate is undefined for \( t \geq t_{(r)} \) as \( n_{r+1} \) is not defined and, for this reason, we have used the original form.

An alternative estimate of \( S(t) \) is from Thomas and Grunkemeier (1975). Their constrained estimate of the survivor function is constructed subject to the condition that it takes a particular value at a specified time. This makes it suitable for use when conducting a significance test as at time \( t^* \) we will have a particular hypothesised value, \( p^* \) of \( S(t^*) \). The constrained estimate of \( S(t) \) is

\[
\tilde{S}(t; t^*, p^*) = \prod_{j:t_{(j)} \leq t} \frac{n_j + \lambda - d_j}{n_j + \lambda}, \quad 0 \leq t \leq t^*, \quad (1)
\]

where \( \lambda \) is chosen such that \( \tilde{S}(t^*; t^*, p^*) = p^* \). This value of \( \lambda \) can be found by a simple numerical search. Just as \( \hat{S}(t) \) is used to derive Greenwood’s formula, we can derive the constrained estimate of variance using \( \hat{S}(t; t^*, p^*) \), giving

\[
\hat{\sigma}_C^2(t^*; p^*) = (p^*)^2 \sum_{j:t_{(j)} \leq t^*} \frac{\tilde{S}(t_{(j)}; t^*, p^*)}{n_j \hat{S}(t_{(j)}; t^*, p^*)} \tilde{q}_j
\]

\[\]
where $\tilde{p}_j = (n_j + \lambda - d_j)/(n_j + \lambda)$, $\tilde{q}_j = 1 - \tilde{p}_j$ and $\tilde{S}(t)$ and $\tilde{S}(t; t^*, p^*)$ denote values of $S(t)$ and $S(t; t^*, p^*)$ just before time $t$. This estimate is used by Thomas and Grunkemeier for a likelihood ratio test (see §2.3) and by Jennison and Turnbull (1985) to directly test $H_0$. Note that the formula stated by Jennison and Turnbull (p.620) is in error as, in their notation, it has $\tilde{S}(T_j)$ and $\tilde{S}(T_j)$ rather than $\tilde{S}(T_j)$ and $\tilde{S}(T_j)$.

The test used by Rothman (1978) in constructing a confidence interval for a survival probability also relies on the approximate normality of $\tilde{S}(t^*)$. Rothman’s estimate of $\sigma^2(t^*)$ is derived by reference to the uncensored case where $\text{var}\{\tilde{S}(t^*)\} \propto S(t^*)\{1 - S(t^*)\}$ and Greenwood’s formula reduces to $\sigma^2_G(t^*) = \tilde{S}(t^*)\{1 - \tilde{S}(t^*)\}/n$, equalling the exact binomial variance if $\tilde{S}(t^*) = S(t^*)$. When testing $H_0: S(t^*) = p^*$ for censored data, Rothman approximates $\text{var}\{\tilde{S}(t^*)\}$ by

$$\hat{\sigma}^2_R(t^*; p^*) = \frac{p^*(1 - p^*)}{S(t^*)\{1 - S(t^*)\}} \hat{\sigma}^2_G(t^*).$$

This matches the binomial variance for uncensored data in that it is proportional to $p^*(1 - p^*)$ and agrees with Greenwood’s formula if $\tilde{S}(t^*) = p^*$. The consistency of $\tilde{S}(t^*)$ (Meier, 1975) ensures that $\hat{\sigma}^2_R(t^*; p^*)$ and $\hat{\sigma}^2_G$ are asymptotically equivalent when $S(t^*) = p^*$. This estimate of $\text{var}\{\tilde{S}(t^*)\}$ can be regarded as a binomial variance with “effective sample size”

$$n' = \tilde{S}(t^*)\{1 - \tilde{S}(t^*)\}/\hat{\sigma}^2_G.$$

These variance estimates will all be used in the methods described in §2.1 for testing $H_0: S(t^*) = p^*$. Some other methods which have been proposed are outlined in §2.2 – 2.4.

## 2 Testing $H_0: S(t^*) = p^*$

### 2.1 Tests using the asymptotic normality of $\tilde{S}(t)$

The result from Meier (1975) that as $n \to \infty$,

$$\frac{\tilde{S}(t) - S(t)}{\sqrt{\text{Var} \{\tilde{S}(t)\}}} \to N(0, 1)$$

3
in distribution for any consistent estimator of \( \text{var}\{\hat{S}(t)\} \) can be used to test the null hypothesis \( H_0: S(t^*) = p^* \). If \( \text{var}\{\hat{S}(t^*)\} \) is consistently estimated by \( \hat{\sigma}^2(t^*) \), then \( H_0 \) will be rejected at the 100\( \alpha \)% level if \( \hat{S}(t^*) \not\in \{p^* \pm z_{\alpha/2} \hat{\sigma}(t^*)\} \), where \( z_{\alpha/2} \) is the upper 100\( (1 - \alpha/2) \)% point of the standard normal distribution. The estimate of \( \text{var}\{\hat{S}(t^*)\} \) can be \( \hat{\sigma}^2_G(t^*), \hat{\sigma}^2_P(t^*), \hat{\sigma}^2_R(t^*; p^*) \) or \( \hat{\sigma}^2_C(t^*; p^*) \).

Figure 1 shows empirical type I error rates for the Greenwood test in 5,000 simulated data sets. In each case, sample size \( n = 100 \) and failure times followed an exponential distribution with mean 10, denoted \( \text{Exp}(10) \). Censoring times were from an \( \text{Exp}(50) \) distribution, so the probability of an observation being censored was 1/6. All subsequent results reported in this section and \( \S 2.2 - \S 2.4 \) are for the same \( n \), censoring and failure distributions unless noted otherwise.

In Figure 1 and all subsequent similar figures the left bar of each pair represents rejection of \( H_0 \) due to a low value of \( \hat{S}(t^*) \) and the right bar rejection of \( H_0 \) due to a high value of \( \hat{S}(t^*) \). There are 17 pairs of bars for \( p^* \) running from 0.1 to 0.9 in steps of 0.05 and \( t^* \) chosen such that \( S(t^*) = p^* \) in each case. The intended two-sided type I error is \( \alpha = 0.05 \) and the central horizontal line at \( \alpha/2 = 0.025 \) indicates the target height for each bar. The two dotted lines...
Peto test

Figure 2: Peto test error rates with $n = 100$, survival times $\sim \text{Exp}(10)$ and censoring times $\sim \text{Exp}(50)$

at $\alpha = 0.025 \pm 0.006$ denote the range within which the observed one-sided error rate from a sample of 5,000 should lie with probability 0.95 if the true one-sided error probability is 0.025. Figure 1 demonstrates that the Greenwood test encounters serious difficulties for low or high values of $S(t)$. Overall error rates are too high and the tails are severely and systematically asymmetric.

Figure 2 shows empirical type I error rates for the “Peto test” which uses Peto’s estimate of variance in testing $H_0$. As for the Greenwood test, results are poorest for low or high survival probabilities, and errors in the two tails are very asymmetric. There is no sign of an improvement over the Greenwood test and results for low $S(t^*)$ are noticeably inferior, despite this being a case where $\sigma^2_P(t^*)$ was intended to improve on $\sigma^2_G(t^*)$.

Figure 3 shows empirical type I error rates for the “constrained variance” test which uses the constrained estimate of variance in testing $H_0$. The results are far better than those seen in Figs 1 and 2 over the whole range of values of $S(t^*)$. There is evidence of asymmetry between the two tails for $S(t^*)$ near 0 or 1 but it is now much less severe.

Note that if $\tilde{S}(t^*) = 1$, $\tilde{S}(t; t^*, p^*)$ is undefined and we test $H_0$ directly but conservatively.
Let $n_c$ denote the number of observations censored before $t^*$. Placing the failure probability $1 - p^*$ just before $t^*$, the number of subjects surviving past $t^*$, conditional on $n_c$ being censored before $t^*$, is $\text{Bin} (n - n_c, p^*)$. In the observed data all $n - n_c$ uncensored cases survived past $t^*$, and we reject $H_0$ if $(p^*)^{(n-n_c)} \leq \alpha/2$.

Figure 4 shows empirical type I error rates for the Rothman test. These results are similar to those for the constrained variance test in Figure 3, although slightly worse for low values of $S(t^*)$. Other simulation studies, not reported here, also support the conclusion that the constrained variance and Rothman tests attain type I error rates closer to their target values over the range of $S(t^*)$ than do the Greenwood and Peto tests. We have also used the variance estimate $\sigma^2_P(t^*)$ in place of $\sigma^2_G(t^*)$ in Rothman’s estimate of $\text{var}\{\hat{S}(t^*)\}$ but, despite improvements in some instances, results were somewhat inferior overall.

To gain some insight as to the relative performance of the methods, and why the results are so much worse for low or high $S(t)$, it is necessary to look at the empirical distributions of $\hat{S}(t)$ and the variance estimates.
Rothman test

![Rothman test error rates graph]

Figure 4: Rothman test error rates with $n = 100$, survival times $\sim \text{Exp}(10)$ and censoring times $\sim \text{Exp}(50)$

2.1.1 Sampling distribution of $\hat{S}(t)$

To examine the distribution of $\hat{S}(t)$ empirically, the histograms in Figure 5 were plotted. These histograms are of samples of $\hat{S}(t)$ for $S(t) = 0.1, 0.5$ and $0.9$ respectively. For all these samples, the failure times were exponentially distributed with mean 10 and the censor times were from an $\text{Exp}(25)$ distribution. In each case, 5,000 samples were taken, but in Figure 5A, 24 samples were undefined for $\hat{S}(23.03)$, as the last observation was a censoring time and before $t = 23.03$, the time such that $S(t) = 0.1$. These cases were omitted from this figure. Note that these cases represent 0.5% of the samples — a value which seems small, but is significant in comparison to the target error rates of 2.5%. This is another reason for caution in cases where survival probability is clearly low. This problem increases with decreasing $n$ and increased censoring.

On these histograms, the overlaid line is the density function of a normal distribution with mean $S(t)$ and variance the sample variance of the values of $\hat{S}(t)$. For Figure 5A, the sample variance was of the 4,976 values of $\hat{S}(23.03)$ which were defined.

Figure 5B shows that near the median, the normal approximation is a good one, and Figure 5A and C show that even for low or high $S(t)$ the approximation seems to be at least somewhat reasonable although there is some skewing of the sample for low or high $S(t)$. This skewing increases as $n$ decreases and the relative degree of censoring increases. Also noteworthy is the
Figure 5: Histograms of samples of estimates $\hat{S}(t)$ when $S(t) = 0.1$, 0.5 and 0.9. Sample size $n$ is 100, survival times are from $\text{Exp}(10)$ and censoring times from $\text{Exp}(25)$. Overlaid curves are normal density functions with mean $S(t)$ and variance the sample variance of the simulated values of $\hat{S}(t)$.
Figure 6: 4,976 estimates of $S(t)$ when $S(t) = 0.1$, with normal distributions based on Greenwood’s estimate of variance. The curves have mean 0.1, and variance equal to the sample variance of the simulated $\hat{S}(t)$, smallest non-zero variance and greatest variance produced by Greenwood’s formula respectively.

bar at zero in Figure 5A. The values of $\hat{S}(t)$ represented by this bar are all zero, and are the result of samples where the last observation was a failure time and occurred before $t = 23.03$. Again, this deviation from normality increases for lower $n$ and increased censoring.

However, each separate confidence interval uses a different value for the variance, and the fitted distributions thus vary greatly. For example, Figure 6 shows the same sample as figure 5A, with the same normal distribution overlaid. This figure also shows two other normal distributions, one with variance equal to the largest estimated variance (using Greenwood’s formula) and one with the smallest non-zero variance produced by Greenwood’s formula.

Clearly, this variation is troubling, as the normal distributions which are being implicitly fitted by the intervals are thus very variable. Similar plots for the variances estimated by Peto’s formula and the constrained method are shown in Figures 7 and 8. These show that the intervals using Greenwood’s variance estimate are slightly more stable than those using Peto’s
Figure 7: 4,976 estimates of $S(t)$ when $S(t) = 0.1$, with normal distributions based on Peto’s estimate of variance. The curves have mean $0.1$, and variance equal to the sample variance of the simulated $S(t)$, smallest non-zero variance and greatest variance produced by Peto’s formula respectively.

variance estimate. They also show that the intervals using the constrained estimate of variance are much less variable. This is one reason for the relative performance of these methods.

Similar plots for the case when $S(t) = 0.5$ show little variation between the fitted distributions, as would be expected from the performance of the intervals in this case. For $S(t) = 0.9$, a similar degree of variability to that shown in figures 6 to 8 is seen, but with less difference between the intervals using Greenwood’s and Peto’s variance estimates.

One other element of these empirical distributions is worth noting and can be seen in Figure 9. This is a histogram of the same sample of $\hat{S}(t)$ when $S(t) = 0.9$ as in figure 5C, and shows a very discrete pattern due to the form of the Kaplan–Meier estimator. The main element of $\hat{S}(t)$ at any given $t$ is the number of failures, with some effect from the number of censored observations to that time. Here there has been little censoring to “blur” the pattern, and the value taken by $\hat{S}(t)$ is almost precisely determined by the number of failures.
Figure 8: 4976 estimates of $S(t)$ when $S(t) = 0.1$, with normal distributions based on the constrained estimate of variance. The curves have mean 0.1, and variance equal to the sample variance of the simulated $\hat{S}(t)$, smallest variance and greatest variance estimated by the constrained method respectively.

2.1.2 Variance estimator distributions

From the simulation studies in §2.1, it is clear that intervals using both Greenwood’s and Peto’s formulae have problems for low and high survival probabilities. Two partial causes for this have already been suggested, and a further cause is shown in Figures 10 and 11.

These plots are of the values Greenwood’s and Peto’s formulae produce as standard error estimates against the corresponding values of $\hat{S}(t)$ when $S(t)$ is 0.1 and 0.9. Clearly, there is an association between $\hat{S}(t)$ and its estimated standard error, $\hat{se}\left\{\hat{S}(t)\right\}$. This results in unbalanced intervals, as seen in the graphs of error rates for these methods where each pair of bars was consistently unbalanced when $S(t)$ was high or low. For example, when $S(t) = 0.1$, the association between $\hat{S}(t)$ and $\hat{se}\left\{\hat{S}(t)\right\}$ leads to a low estimate of $\hat{S}(t)$ having a low corresponding estimate of $\hat{se}\left\{\hat{S}(t)\right\}$. This means a confidence interval centred below $S(t)$ will tend to be narrower than one centred above $S(t)$. This contributes to the asymmetry of the error rates, with more frequent rejection of $H_0$ for $S(t^*)$ being below the acceptance region than above. The opposite occurs for high $S(t)$, leading to asymmetry in the opposite direction.
In particular, note that if $\hat{S}(t) = 0$, using either Greenwood's or Peto's formulae gives $\text{se}\left\{\hat{S}(t)\right\} = 0$, resulting in a degenerate confidence interval of $[0]$ for $S(t)$. Equivalently, this means a hypothesis $H_0: S(t) = p$ will be rejected for any $p \neq 0$. Similarly, if $\hat{S}(t) = 1$ then both these variance estimates give $\text{se}\left\{\hat{S}(t)\right\} = 0$, resulting in a confidence interval of $[1]$ for $S(t)$ and rejection of $H_0$ for any $p \neq 1$.

In contrast, similar plots for $S(t) = 0.5$ show much less association between $\hat{S}(t)$ and the estimated standard error, as reflected in the better performance in that region. Similar plots for the constrained estimate of standard error show less of an association as the assumption of $S(t^*) = p^*$ has a stabilising effect on $\hat{\sigma}_C^2(t^*)$. These plots are in Figure 12. Similarly, equivalent plots of Rothmans’s estimate of standard error in Figure 13 show the same lack of dependence on the value taken by $\hat{S}(t^*)$. Also, the constrained estimate will give $\hat{\sigma}_C(t^*) > 0$ even if $\hat{S}(t) = 0$, thus eliminating the problem of degenerate confidence intervals in this case. As noted previously, if $\hat{S}(t^*) = 1$, then we test $H_0$ directly, thus avoiding the problem of degenerate confidence intervals.

This association can be seen clearly in the case of no censoring, when Greenwood’s variance
Figure 10: Greenwood’s estimates of $\text{se}\{\hat{S}(t)\}$ plotted against corresponding values of $\hat{S}(t)$ when $S(t) = 0.1$ (left plot) and when $S(t) = 0.9$ (right plot).

Figure 11: Peto’s estimates of $\text{se}\{\hat{S}(t)\}$ plotted against corresponding values of $\hat{S}(t)$ when $S(t) = 0.1$ (left plot) and when $S(t) = 0.9$ (right plot).
Figure 12: *Constrained estimates of se\{\hat{S}(t)\} plotted against corresponding values of \(\hat{S}(t)\) when \(S(t) = 0.1\) (left plot) and when \(S(t) = 0.9\) (right plot).

Figure 13: *Rothman’s estimates of se\{\hat{S}(t)\} plotted against corresponding values of \(\hat{S}(t)\) when \(S(t) = 0.1\) (left plot) and when \(S(t) = 0.9\) (right plot).
estimate reduces to \( \hat{S}(t)(1 - \hat{S}(t))/n \) and the original form of Peto’s estimate reduces to
\[
\hat{S}(t)(1 - \hat{S}(t))\frac{n_{k+1}}{m_k} - \text{both of these depend on } \hat{S}(t).
\]
In contrast, when there is no censoring the constrained estimate reduces to \( S(t)(1 - S(t))/n \), a constant expression.

Note also that these plots all show some degree of discretisation, or clustering. It is greater for those plots where \( S(t) = 0.9 \), as with the earlier histogram of \( \hat{S}(t) \) values, figure 9. As in this figure, the discretisation is due to the dependence on the number of failures at time \( t \), which is obviously a discrete quantity. The plots of Peto’s estimates of standard error also feature this discretisation strongly, largely due to the fact that unlike the other methods, Peto’s formula is simply one term using \( \hat{S}(t) \) and \( n_k \), while the other methods use a summation over failure times, which has the effect of blurring the discretisation.

2.2 An improvement on the normal approximation to the distribution of \( \hat{S}(t^*) \)

Since \( \hat{\sigma}_C^2(t^*; p^*) \) and \( \hat{\sigma}_R^2(t^*; p^*) \) perform so well as estimators of \( \text{var}\{\hat{S}(t^*)\} \), the most plausible reason for shortcomings of the constrained variance and Rothman tests, when they occur, is inadequacy of the normal approximation to the distribution of \( \hat{S}(t^*) \). Anderson, Bernstein and Pike (1982) suggest applying logit, arcsin or \( \log(-\log) \) transformations to \( \hat{S}(t^*) \) and Collett (Collett, 1994) suggests the probit transformation to produce a variable which is more nearly normally distributed. We have found the \( \log(-\log) \) transformation to gives the best results. In this case \( \log[-\log\{\hat{S}(t^*)\}] \) is treated as approximately normal with estimated variance
\[
\hat{\sigma}_G^2(t^*)/[\hat{S}(t^*) \log\{\hat{S}(t^*)\}]^2 \quad \text{and} \quad H_0: S(t^*) = p^* \text{ is rejected if}
\]
\[
\log[-\log\{\hat{S}(t^*)\}] \not\in \left\{ \log\{-\log(p^*)\} \pm z_{\alpha/2} \hat{\sigma}_G(t^*)/[\hat{S}(t^*) \log\{\hat{S}(t^*)\}] \right\}. \tag{2}
\]

Empirical type I error rates for this “transformed test” are shown in Figure 14. Replacing \( \hat{\sigma}_G(t^*) \) in (2) by \( \hat{\sigma}_P(t^*) \) or \( \hat{\sigma}_C(t^*; p^*) \) led to worse results. In these and other simulations we have found the transformed test to be inferior to the constrained variance and Rothman tests. The test is produces more errors with \( \hat{S}(t^*) \) too low than too high and problems occur when \( \hat{S}(t^*) = 0 \) or \( 1 \), in which case both \( \log[-\log\{\hat{S}(t^*)\}] \) and its estimated variance are undefined.
Figure 14: Transformed test error rates with \( n = 100 \), survival times \( \sim \text{Exp}(10) \) and censoring times \( \sim \text{Exp}(50) \)

We have used the data from our simulations to investigate the distribution of transformed versions of \( \tilde{S}(t^*) \) but have found no satisfactory transformation to normality.

2.3 Thomas and Grunkemeier’s likelihood ratio test

As was previously noted, Thomas & Grunkemeier (1978) derived the constrained estimate of \( S(t), \tilde{S}(t; t^*, p^*) \). Their motivation in this was to use \( \tilde{S}(t; t^*, p^*) \) in constructing a likelihood ratio confidence interval for \( S(t^*) \). The equivalent test of \( H_0: S(t^*) = p^* \) rejects \( H_0 \) at the 100\( \alpha \)% level if

\[
-2R > \chi^2_{1, \alpha}
\]

where \( R \) is the log-likelihood ratio statistic

\[
R = \sum_{i: t_i \leq t^*} \left\{ (n_i - d_i) \log \left( 1 + \frac{\lambda}{n_i - d_i} \right) - n_i \log \left( 1 + \frac{\lambda}{n_i} \right) \right\};
\]

for more details of the derivation of this expression, see the paper by Thomas & Grunkemeier. The authors outline how \(-2R\) may be shown to be asymptotically \( \chi^2_1 \) distributed, and later work by Li (1995) gives a rigorous proof of this.

Figure 15 shows simulation results for Thomas and Grunkemeier’s likelihood ratio method. The method clearly gives good results, in this case performing better than all the other methods seen so far. However, a small degree of systematic asymmetry is still present.
Likelihood ratio test

\[ S(t*) \]

\[ 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0, 0.01, 0.02, 0.03, 0.04, 0.05 \]

Error rates

Figure 15: Likelihood ratio test error rates with \( n = 100 \), survival times \( \sim \) \( \text{Exp}(10) \) and censoring times \( \sim \) \( \text{Exp}(50) \)

2.4 Bootstrap significance tests

The bootstrap offers an alternative to parametric approximations to the distribution of \( \hat{S}(t^*) \). Efron (1981) and Reid (1981) have proposed bootstrap methods for making inferences about quantiles of a survival distribution, although Reid’s method has been shown to be asymptotically conservative when censoring is present (Akritas, 1986).

The endpoints of Efron’s (1981) “percentile” confidence intervals for functions of the survival distribution are simply percentiles of the bootstrap sampling distribution of the quantity of interest. If his \( 100(1 - \alpha)\% \) percentile confidence interval for \( S(t^*) \) or \( S^{-1}(p^*) \) is viewed as the inversion of a family of size \( 1 - \alpha \) hypothesis tests, the corresponding test of \( H_0: S(t^*) = p^* \) rejects \( H_0 \) if \( p^* \) lies in the lower \( \alpha/2 \) or upper \( \alpha/2 \) tail of the bootstrap distribution of \( \hat{S}(t^*) \). Efron’s bootstrap data sets are obtained by sampling with replacement from the set of \( n \) observed pairs of survival times and censoring indicators, \( \{(t_i, \delta_i); i = 1, \ldots, n\} \). The Kaplan-Meier estimators of the survival function at time \( t^* \) for each of \( M \) bootstrap samples are written in increasing order as \( \hat{S}_{[1]}^B(t^*) \leq \cdots \leq \hat{S}_{[M]}^B(t^*) \) and \( H_0: S(t^*) = p^* \) is rejected if

\[ p^* < \hat{S}_{[M\alpha/2]}^B(t^*) \text{ or } p^* > \hat{S}_{[M(1-\alpha/2)]}^B(t^*) . \]

If the random variable \( \hat{S}(t^*) - S(t^*) \) is approximately pivotal, its distribution can be
Figure 16: Error rates for the “percentile” and the “pivotal” bootstrap tests with \( n = 100 \), survival times \( \sim \text{Exp}(10) \) and censoring times \( \sim \text{Exp}(50) \)

approximated by that of the bootstrap quantity \( \tilde{S}^B(t^*) - \tilde{S}(t^*) \). This leads to a different form of bootstrap confidence interval; see Hall (1992, Example 1.2) and Efron & Tibshirani (1993, Section 13.4). In this case, the related test rejects \( H_0: S(t^*) = p^* \) if

\[
p^* < 2\tilde{S}(t^*) - \tilde{S}^B_{[M_{1-\alpha/2}]}(t^*) \quad \text{or} \quad p^* > 2\tilde{S}(t^*) - \tilde{S}^B_{[M_{\alpha/2}]}(t^*).
\]

We shall refer to this as the “pivotal” bootstrap test.

Empirical type I error probabilities for the percentile and pivotal tests, using bootstrap samples of 1,000 replicates, are shown in Figure 16. The results are poor and qualitatively similar to those of the original Greenwood test. We attribute this poor performance to the fact that \( \tilde{S}(t^*) - S(t^*) \) is in truth far from pivotal, its variance and skewness both changing substantially with \( S(t^*) \). Since \( \text{var}\{\tilde{S}(t^*)\} \) has a maximum close to \( t^* = 0.5 \), a variance stabilising transformation of \( \tilde{S}(t^*) \) must stretch out values more the further they are from 0.5 and this automatically counteracts some of the skewness in the distribution of \( \tilde{S}(t^*) \). Thus, the percentile method’s suitability to problems where a transformation to standard normality exists (Efron & Tibshirani, 1993, Section 13.3) explains its modest improvement over the Greenwood test, whereas the manner in which the variance and skewness of \( \tilde{S}(t^*) \) vary with \( t^* \) happens to work against the pivotal test.

Methods of obtaining bootstrap confidence intervals which improve on the coverage properties of the simple percentile and pivotal methods have been developed. Possibilities
include percentile-t intervals (Hall, 1992, Section 3.2) and bias corrected and accelerated (hereafter BCa) percentile intervals (Efron, 1987). We have also calculated coverage rates of percentile-t and BCa intervals for the same examples. Results for the percentile-t method were poor. Results for the BCa intervals improved on those for the simple percentile and pivotal methods reported earlier and are shown in Figure 17.

3 Literature review

Various other methods have been proposed for producing approximate confidence intervals for censored data. Most of the literature looks at finding time intervals for a given probability, especially confidence intervals for the median, i.e., intervals $[t_l, t_u]$ for $t_m = \inf \{ t : S(t) \leq 0.5 \}$. The coverage failure rates for these intervals are equivalent to the type I error rates when using the appropriate method to test $H_0: S(t^*) = p^*$.

Initially, interval estimation was done by Greenwood’s method, following Kaplan and Meier (1958). Peto’s method was an early alternative (Peto et al., 1977).

Efron (1981) and Reid (1981) proposed two different resampling schemes, leading to bootstrap confidence intervals for the median survival time. Efron also uses bootstrap methods
to estimate $\text{var}\left\{ \hat{S}(t) \right\}$, finding close agreement to the Greenwood estimate of variance in an example. In the same paper, Efron uses a bootstrap method in a simulation and finds good coverage in samples as small as 21. Reid uses a different resampling scheme, which has poorer performance than Efron’s in a simulation. Akritas (1986) uses both resampling schemes in other bootstrap methods and finds below nominal coverage for Reid’s scheme but good coverage for methods using Efron’s scheme. More details on these resampling schemes are included in the discussion on bootstrap methods in §2.4.

Both Brookmeyer and Crowley (1982) and Emerson (1982) generalise the sign test to cope with censored data and propose intervals for the median survival time. Brookmeyer and Crowley approximate the binomial distribution of the test statistic by a normal distribution and in doing so find simple Greenwood intervals. They compare these by simulation with two parametric methods designed for exponential data, finding better coverage for the nonparametric method in cases where the failure times are from other distributions than exponential, as would be expected. The intervals are, however, anti-conservative, in the sense that they have uniformly below nominal coverage. Jennison and Turnbull (1985) suggest a reason for this, noting that Greenwood’s estimate of variance underestimates $\text{var}\left\{ \hat{S}(t_m) \right\}$ for $\hat{S}(t_m) \neq 0.5$, where $t_m$ is the median survival time defined above.

In contrast, Emerson uses a binomial distribution in constructing his intervals using essentially the same test statistic. He compares this against the same alternatives as Brookmeyer and Crowley, finding all methods having similar coverage for exponential data. However, in this case, the parametric methods produce uniformly shorter intervals. For data from other distributions, Emerson reports conservative results for various cases. However, Slud, Byar and Green (1984) disagree, finding Emerson’s method to be noticeably anti-conservative. They suggest this is due to incorrect implementation of the method in Emerson’s simulation.

The paper by Slud, Byar and Green compares the methods from Brookmeyer and Crowley, Emerson, Efron, Reid and Simon and Lee with two alternatives they propose, all for intervals for the median survival time. Their alternatives are found to have asymptotically correct coverage and the intervals thus produced are found to converge to these from Efron and Brookmeyer
and Crowley. Emerson’s and Reid’s methods are found to be asymptotically anticonservative. Simulation results find Brookmeyer and Crowley intervals anticonservative in all cases, as noted above, and the methods by Emerson and Reid become anticonservative as censoring increases. The alternatives proposed by the authors are found to be conservative in all cases, and the authors conclude these methods are to be preferred on the grounds that greater than nominal coverage is to be preferred to performance below nominal rates.

In the paper by Jennison and Turnbull (1985), the constrained variance interval for the median is used and shown to have asymptotically correct error rates, and a small simulation shows excellent results, which are symmetric. The same simulation demonstrates again the anticonservatism of the Brookmeyer and Crowley intervals, and also shows that the methods proposed by Slud, Byar and Green are severely asymmetric.

A paper by Zhao (1996) proposes a new estimate of $\text{var}\left\{\hat{S}(t)\right\}$, the homogenetic estimate of variance, which is shown to reduce to binomial variance in the absence of censoring. This variance estimate is defined in terms of grouped data, when observations are collected into time intervals. The paper also discusses the difference between the original form of Peto’s estimate and Simon and Lee’s expression for this estimate, finding that the original does not reduce to binomial variance while Simon and Lee’s form does. For ungrouped data, the homogenetic estimate of variance reduces to Simon and Lee’s form for Peto’s estimate.

A simulation comparison of Greenwood’s method, the homogenetic estimate and Peto’s method using Simon and Lee’s expression is carried out, for times $t$ such that $S(t) = 0.3, 0.5$ and $0.7$. The three methods produce identical results in the absence of censoring. When censoring is present, the coverage rates are lowest for Greenwood’s method and highest for the Simon and Lee intervals.

In most of these papers, no attempt is made to assess the symmetry of the methods proposed — an oversight highlighted by Slud, Byar and Green’s recommendation of methods later shown to be asymmetric. Only Thomas and Grunkemeier, Anderson, Bernstein and Pike and Jennison and Turnbull look at this aspect of the performance of the methods.

Also unfortunate is that so few papers look at the performance of intervals away from the median. While some of the methods are developed specifically for the median, they could be
adapted for other quantiles. Even in those cases where the method could be applied for other times without alteration most of the authors look solely at the performance at the median. Only Thomas and Grunkemeier, Anderson, Bernstein and Pike and Zhao look at alternative cases, while Efron looks at alternative location statistics such as trimmed means.

While the median is obviously an important case, and median survival time an important statistic, it has been seen throughout the methods in §2 that it is also the case which proves easiest to deal with — there are few cases in the simulations shown where performance at the median is unsatisfactory.

4 Conclusions

We have shown that the methods most commonly mentioned in texts on applied survival analysis, Greenwood’s test, Peto’s test and tests based on normal transformations, can have severe shortcomings when a small sample of censored data is used to test whether a survival probability is equal to a specified value, away from 0.5. There are several alternative methods which perform considerably better. Of all the methods we have described, the best results are from Thomas and Grunkemeier’s likelihood ratio test. This conclusion is supported by empirical error rates we have seen in other simulations with other values of $n$, failure distributions and censoring distributions. Even so, this method encounters difficulties when $n$ is lower and a greater degree of censoring is present.

From the remaining methods, the constrained variance and Rothman methods produce the best results. One useful point of these methods is that they provide an estimate of the distribution of $\hat{S}(t^*)$, which the likelihood ratio test does not.

It is noteworthy that the three best methods all use the information contained in the hypothesis $H_0: S(t^*) = p^*$ — the likelihood ratio and constrained variance methods via $\hat{S}(t; t^*, p^*)$ and Rothman’s method via $p^*$ being used in the variance estimate $\hat{\sigma}_R^2(t^*; p^*)$. The stabilising effect of this information is a key in the improvements these methods show over their competitors. The constrained estimate of the survivor function, $\hat{S}(t; t^*, p^*)$, and the constrained variance estimate, $\hat{\sigma}_C^2(t^*; p^*)$, are also used by Barber and Jennison (1998) in two new types
of test. In the first of these tests, the null distribution of $\hat{S}(t^*)$ is approximated by a Beta distribution. The second test uses $\hat{S}(t; t^*, p^*)$ to approximate the null distribution of $\hat{S}(t^*)$ in a “constrained bootstrap” which is particularly successful, matching the performance of Thomas and Grunkemeier’s likelihood ratio test, and providing a more precise, conservative solution to problems caused by discreteness in the distribution of $\hat{S}(t^*)$.

References


