Bootstrap Confidence Intervals for a Hazard Ratio when the Number of Observed Failure is Small, with Applications to Group Sequential Survival Studies

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

http://people.bath.ac.uk/mascj

Statistical Planning of Translational Studies

Göttingen

March, 2024

Reference

I presented this material at the Interface Conference at the University of Michigan in 1990.

My paper

"Bootstrap Confidence Intervals for a Hazard Ratio when the Number of Observed Failure is Small, with Applications to Group Sequential Survival Studies"

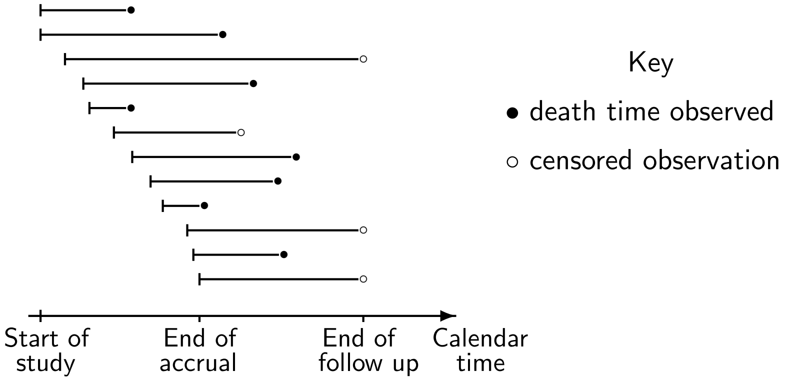
appears in pages 89–97 of

Computing Science and Statistics: Statistics of Many Parameters: Curves, Images, Spatial Models (1992), eds C Page and R LePage.

If a large survival trial is conducted group sequentially, one may expect small numbers of failures at early analyses, so small sample methods become important.

Conducting a trial with a survival endpoint

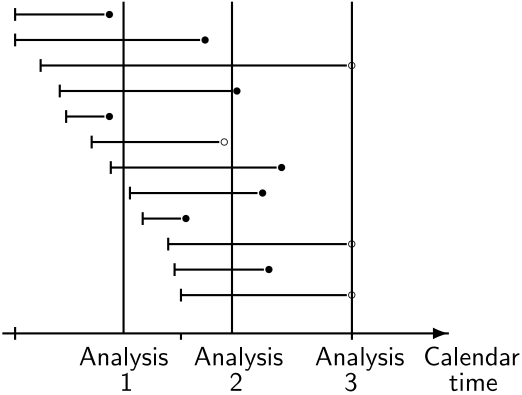
Consider a Phase III trial comparing a new treatment and a control.



Subjects are randomised to a treatment as they enter the study.

Survival is measured from entry to the study.

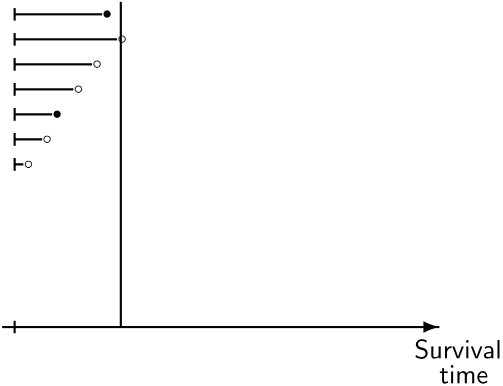
Interim analyses



At an interim analysis, subjects are censored if they are still alive.

Information on such patients continues to accrue at later analyses.

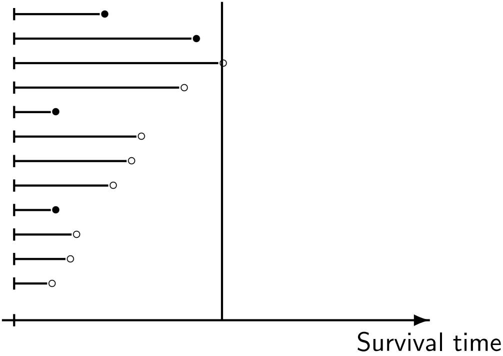
Interim analysis 1



We analyse data on survival from time of randomisation.

Survival times start at zero and "analysis time" censoring occurs for subjects surviving past this first analysis.

Interim analysis 2



At interim analysis 2, there is further follow-up of subjects who were censored at analysis 1.

In addition, there is initial information on the survival times of subjects entering the trial since analysis 1.

The proportional hazards model for survival data

The hazard rate at time t is defined as

$$h(t) \; = \; \lim_{\delta t \to 0} \, \frac{1}{\delta t} \, Pr\{\text{Fail in } [t, t+\delta t) \mid \text{ Survive up to time } t\}.$$

In the proportional hazards model

Treatment A: hazard rate $= h(t)$

Treatment B: hazard rate $= \lambda \, h(t)$

We aim to test sequentially H_0: $\lambda = \lambda_0$ against $\lambda \neq \lambda_0$, with type I error probability $\alpha/2$ in each tail.

This could be

A simple test of H_0: $\lambda = \lambda_0$,

To construct a sequence of Repeated Confidence Intervals for λ,

To compute a Confidence Interval for λ after a sequential test.

The logrank statistic for testing H_0: $\lambda = 1$

At stage k, the observed number of deaths is d_k.

Elapsed times between entry to the study and these deaths are

$$\tau_{1,k} < \tau_{2,k} < \ldots < \tau_{d_k,k} \quad \text{(assuming no ties).}$$

Define variables at analysis k

$r_{iA,k}$ and $r_{iB,k}$	Numbers at risk on Trts A and B at $\tau_{i,k}-$
$r_{ik} = r_{iA,k} + r_{iB,k}$	Total number at risk at $\tau_{i,k}-$
O_k	Observed number of deaths on Trt B
$E_k = \sum_{i=1}^{d_k} r_{iB,k}/r_{ik}$	"Expected" number of deaths on Trt B
$V_k = \sum_1^{d_k} r_{iA,k} r_{iB,k}/r_{ik}^2$	"Variance" of O_k
$Z_k = (O_k - E_k)/\sqrt{V_k}$	Standardised logrank statistic

The score statistic for testing H_0: $\lambda = \lambda_0$

This generalisation of the logrank statistic is obtained by differentiating the logarithm of the partial likelihood, as defined by Cox (*Biometrika*, 1975).

The (unstandardised) score statistic at analysis k is

$$L_k(\lambda_0) \; = \; \sum_{i=1}^{d_k} \left(\delta_{i,k} - \frac{\lambda_0 \, r_{iB,k}}{r_{iA,k} + \lambda_0 \, r_{iB,k}} \right)$$

where $\delta_{i,k}$ is the indicator that failure i at analysis k is on Treatment B.

Thus

$$L_k(\lambda_0) \;\; = \;\; \text{Observed number of failures on Treatment B}$$

$$- \text{ "Expected" number of failures if } \lambda = \lambda_0.$$

Large sample distribution of $L_k(\lambda_0)$

Define the information for λ_0 at analysis k as

$$\mathcal{I}_k \ = \ \sum_{i=1}^{d_k} \ \frac{\lambda_0 \, r_{iA,k} \, r_{iB,k}}{(r_{iA,k} + \lambda_0 \, r_{iB,k})^2}.$$

Then asymptotically, if $\lambda = \lambda_0$,

$$\frac{L_k(\lambda_0)}{\sqrt{\mathcal{I}_k}} \ \xrightarrow{\mathcal{D}} \ N(0,1)$$

as the number of observations and the number of observed failures at analysis k tend to infinity.

Furthermore, the asymptotic joint distribution of the sequence $\{L_1, \ldots, L_K\}$ is multivariate normal with independent increments: see Jennison & Turnbull (*JASA*, 1997) and references therein.

An error spending sequential test of H_0: $\lambda = \lambda_0$

With a maximum of K analyses, specify $\pi_1, \ldots, \pi_K$ where

$$\sum_{k=1}^{K} \pi_k \;=\; \alpha.$$

Here π_k represents the error probability "spent" at analysis k.

Compute $c_1, \ldots, c_K$ such that

$$Pr_{\lambda=\lambda_0}\{|L_1(\lambda_0)| < c_1, \ldots, |L_{k-1}(\lambda_0)| < c_{k-1},\, |L_k(\lambda_0)| \geq c_k\} \;=\; \pi_k,$$

assuming the asymptotic normal distribution of $\{L_1, \ldots, L_K\}$.

Then, reject H_0: $\lambda = \lambda_0$ at analysis k if $|L_1(\lambda_0)| \geq c_k$.

The values of $\pi_1, \ldots, \pi_K$ may be fixed in advance (Slud & Wei, *JASA*, 1982) or functions of the observed information, $\mathcal{I}_1, \ldots, \mathcal{I}_K$ (Lan & DeMets, *Biometrika*, 1983).

Accuracy of the normal approximation

In a non-sequential test

In a single sample test, the normal approximation for $L_k(\lambda_0)$ is accurate when the number of failures is large.

It is less accurate if the number of failures is as low as 20 or 30.

With few failures, the distribution of $L_k(\lambda_0)$ is skew for $\lambda_0 \neq 1$.

In a group sequential test

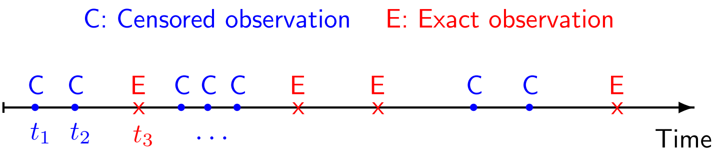
We find error rates in a group sequential test are accurate if the normal approximation to the distribution of each $L_k(\lambda_0)$ is good.

Experience with other response variables suggests that, when the numbers of failures are low, it will suffice to improve the accuracy of the marginal distribution for each $L_k(\lambda_0)$.

Then we shall reject H_0: $\lambda = \lambda_0$ at analysis k if H_0 is rejected in a two-sided test with significance level $2\{1 - \Phi(c_k/\sqrt{\mathcal{I}_k})\}$.

Small sample approximation

We approximate the conditional distribution of $L(\lambda_0)$ given the order of exact and censored survival times.



Generate group membership for events at times t_1, t_2, ... in order.

Start with n_1 in group A and n_2 in group B.

If the next event is **censored**

With probability $n_1/(n_1 + n_2)$,

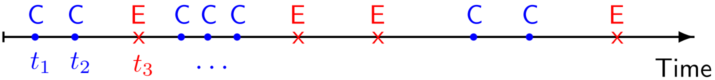
allocate the event to group A, reduce n_1 by 1.

With probability $n_2/(n_1 + n_2)$,

allocate the event to group B, reduce n_2 by 1.

Small sample approximation

C: Censored observation E: Exact observation



If the next event is **exact**

With probability $n_1/(n_1 + \lambda_0\, n_2)$,

allocate the event to group A, reduce n_1 by 1.

With probability $\lambda_0\, n_2/(n_1 + \lambda_0\, n_2)$,

allocate the event to group B, reduce n_2 by 1.

After allocating all the events, evaluate

$$L_k(\lambda_0) \;=\; \sum_{i=1}^{d_k} \left(\delta_{i,k} - \frac{\lambda_0\, r_{iB,k}}{r_{iA,k} + \lambda_0\, r_{iB,k}} \right).$$

Small sample approximation: Validity & implementation

Allocations of events to treatment groups follow the proportional hazards model exactly if there is no censoring or if $\lambda = 1$.

The scheme produces the correct asymptotic distribution as the sample size and number of events tend to infinity.

We can use Monte Carlo or "bootstrap" sampling to test H_0.

For a 2-sided, level α test of H_0: $\lambda = \lambda_0$:

Generate $N - 1$ "bootstrap" values of $L(\lambda_0)$.

If the observed value is one of the $N\alpha/2$ smallest or $N\alpha/2$ largest values in the set of N observations, reject H_0.

If the bootstrap is sampling the correct distribution, the error rate of this procedure is exactly α.

To minimise simulation noise, a very large of N should be used.

Calculating a $100(1-\alpha)\%$ confidence interval for λ

Use the normal approximation to find initial estimates of the endpoints of the confidence interval for λ.

Let

$$p(\lambda) \;=\; Pr\{\text{Bootstrap } L(\lambda) > \text{observed } L(\lambda)\}.$$

Simulate under values of λ in the neighbourhood of each endpoint and model the function $p(\lambda)$ in these regions, e.g., by logistic regression.

Solve the equations

$$p(\lambda) \;=\; \alpha/2$$

and

$$p(\lambda) \;=\; 1-\alpha/2$$

to find the endpoints of the confidence interval for λ.

Assessing the small sample approximation

We shall simulate M data sets under $\lambda = \lambda_0$.

For each data set, we generate N bootstrap samples and use these to decide whether or not to reject H_0: $\lambda = \lambda_0$.

We can compare the error rate in these M simulated data sets to the target value α.

In 1990, I aspired to simulate $M = 20,000$ data sets, giving an estimate of a type I error rate ≈ 0.05 with standard error 0.0015.

These days, I would expect to use $M = 1,000,000$, to give an estimated error rate with standard error 0.0002.

With a high value of M and $N = 1,000$, say, the computation time is considerable.

However, for each data set, we only need to know whether or not H_0 is rejected.

Curtailing the bootstrap test

Let X be the number of bootstrap simulations giving a value of $L(\lambda_0)$ greater than our observed value, $L^*(\lambda_0)$.

With $c = N\alpha/2$, we reject H_0 if

$$X < c \quad \text{or} \quad X > N-1-c \tag{1}$$

and we accept H_0 if

$$c \leq X \leq N-1-c. \tag{2}$$

Deterministic curtailment

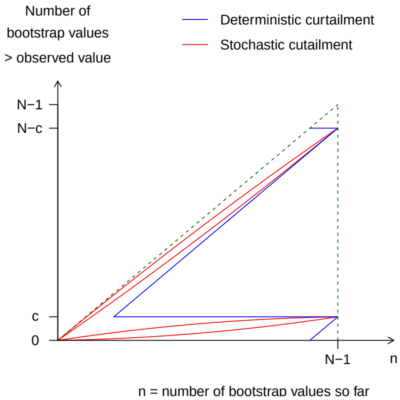
If we already have c bootstrap values greater than $L^*(\lambda_0)$ and c less than $L^*(\lambda_0)$, we know X will satisfy (2), so we can stop now.

Stochastic curtailment

We can stop if the final decision is **almost** inevitable given the bootstrap results so far.

I allowed a maximum probability of 10^{-5} to make an error here.

Curtailing the bootstrap test



Average number of bootstraps required

	$N = 100$	$N = 1,000$	$N = 10,000$
$\alpha/2 = 0.05$	29	82	688
$\alpha/2 = 0.01$	9	84	295

Results: Single sample test

Survival times $\sim$ Exponential with median ≈ 1.

Censoring times $\sim$ Uniform$(0, 1)$.

60 observations, average number of failures $= 17$.

Empirical error rates for test of H_0: $\lambda = \lambda_0$.

		$\alpha/2 = 0.05$		$\alpha/2 = 0.01$	
		$\lambda > \lambda_0$	$\lambda < \lambda_0$	$\lambda > \lambda_0$	$\lambda < \lambda_0$
$\lambda_0 = 2$	Normal approx.	0.045	0.056	0.0067	0.0134
	Bootstrap	0.050	0.050	0.0094	0.0101
$\lambda_0 = 3$	Normal approx.	0.042	0.061	0.0055	0.0144
	Bootstrap	0.049	0.052	0.0111	0.0094
	Standard error	0.0015		0.0007	

Based on $M = 20,000$ replicates, $N = 1,000$ bootstrap samples.

Results: Group sequential test

5 year study, accrual for 2 years then follow-up, 10 analyses.

Median survival $\sim$ 2.5 years.

Target error rate $\alpha/2 = 0.0$.

Average failures at analyses 1, 2, 3, ... = 3.5, 13, 17 ... .

Empirical error rates for test of H_0: $\lambda = \lambda_0$.

		$\lambda > \lambda_0$	$\lambda < \lambda_0$
$\lambda_0 = 2$	Normal approx.	0.036	0.060
	Bootstrap	0.049	0.051
$\lambda_0 = 3$	Normal approx.	0.032	0.067
	Bootstrap	0.049	0.051
	Standard error	0.0015	

Based on $M = 20,000$ replicates, $N = 1,000$ bootstrap samples.

Conclusions

The normal approximation for logrank statistics can be poor when the number of failures is small.

Our small sample approximation is effective and can be used both in fixed sample and group sequential tests.

The bootstrap tests are based on an accurate method for simulating under a hypothesised parameter value.

Stochastic curtailment of bootstrap tests can reduce computation time by a factor as high as 30, making a proper assessment of error rates feasible.

Bootstrap hypothesis tests are generally applicable.

See, for example, Barber & Jennison (*Biometrics*, 1999), "Symmetric tests and confidence intervals for survival probabilities and quantiles of censored survival data".