Comparing Two Kaplan-Meier Curves with the Probability of Agreement

NATHANIEL T. STEVENS

FALL TECHNICAL CONFERENCE 2018

UNIVERSITY OF SAN FRANCISCO

ACKNOWLEDGEMENTS

- Lu Lu, University of South Florida

OUTLINE

- Problem description.
- What is the probability of agreement?
- How can we use this to compare reliability curves?
  - The Asymptotic approach.
  - The Bootstrapping approaches.
- Illustrative Example.
- Conclusions.

Comparing Lifetime Distributions

Comparing lifetime distributions for different populations is important in a variety of fields:

- Reliability Engineering
- Survival Analysis
- Customer Analytics

Common goal:

Compare two lifetime distributions to evaluate their similarity

Contextual goal: compare reliability curves

Comparing Lifetime Distributions

Many parametric and nonparametric methods exist for estimating reliability curves.

We focus on the Kaplan-Meier (KM) estimator[1] which provides a nonparametric estimate of reliability, defined here to be the probability that a product's lifetime exceeds t time units:

$$S(t) = P(T \geq t)$$

Given observed (censored and uncensored) event times $t_1 < t_2 < \cdots < t_r$,

$$\hat{S}(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i}$$

where n_i is the number of units at risk of failing at time point t_i and d_i is the number of units that fail at time point t_i

Comparing Lifetime Distributions

Our goal is to compare $S_1(t)$ and $S_2(t)$

This is commonly done with tests of the following hypothesis:

$$H_0: S_1(t) = S_2(t) \text{ vs. } H_A: S_1(t) \neq S_2(t)$$

- Mantel-Haenszel logrank test[2]
- Gehan-Wilcoxon test[3]
- G^ρ class of tests[4]

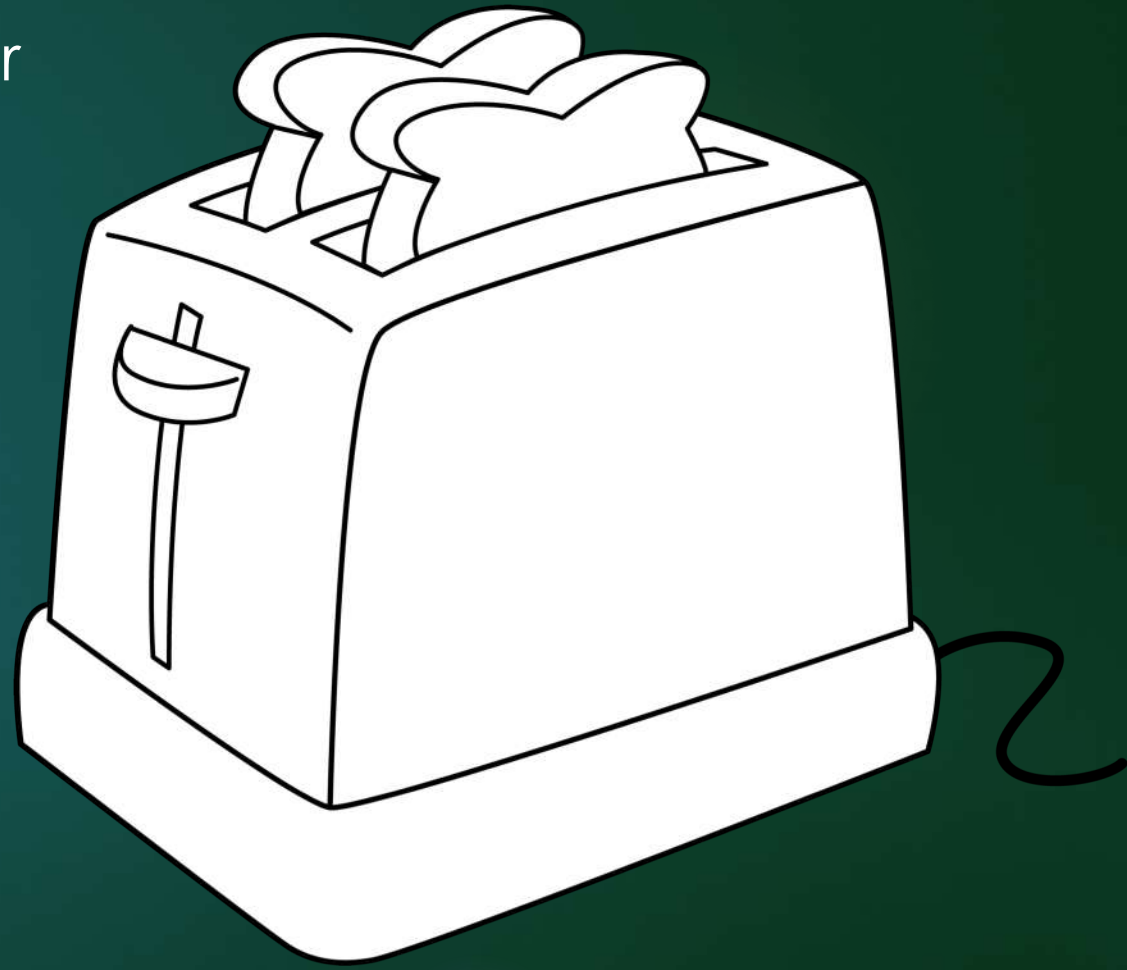
As an alternative we propose to compare $S_1(t)$ and $S_2(t)$ via their KM estimators $\hat{S}_1(t)$ and $\hat{S}_2(t)$ using the probability of agreement

The Toaster Snubber Example

Nelson[5] describes an accelerated life test in which two different versions of a toaster are repeatedly cycled.

The different versions correspond to old and new snubbers (toaster component).

There were $r_1 = 52$ "old" toasters and $r_2 = 54$ "new" toasters involved in this comparison.

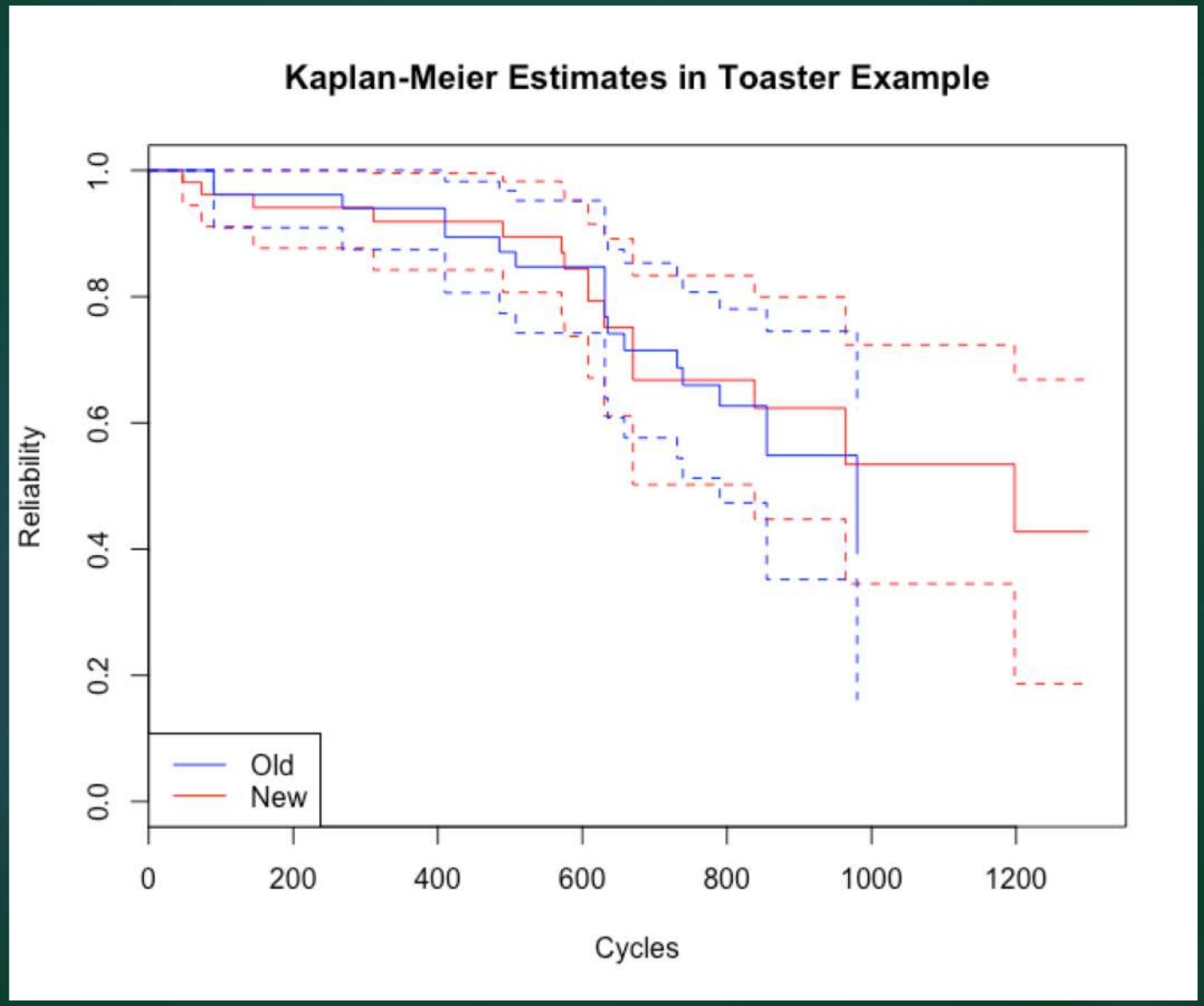


The Toaster Snubber Example

OLD				
90	410	658	790+	980+
90	410+	658+	790+	980+
90+	485	731	790+	980+
190+	508	739	790+	980+
218+	600+	739+	790+	
218+	600+	739+	790+	
241+	600+	739+	790+	
268	600+	739+	790+	
349+	631	790	855	
378+	631	790+	980	
378+	631	790+	980	
410	635	790+	980+	

NEW				
45+	311	608+	670	1164+
47	417+	608+	670	1198
73	485+	608+	731+	1198+
136+	485+	608+	838	1300+
136+	490	608+	964	1300+
136+	569+	608+	964	1300+
136+	571	608+	1164+	
136+	571+	608+	1164+	
145	575	608+	1164+	
190+	608	608+	1164+	
190+	608	608+	1164+	
281+	608+	630	1164+	

The Toaster Snubber Example



Comparing Lifetime Distributions

What do the formal tests say about

$$H_0: S_1(t) = S_2(t) \text{ vs. } H_A: S_1(t) \neq S_2(t) \text{ ?}$$

- Mantel-Haenszel logrank test: *p*-value = 0.697
- Gehan-Wilcoxon test: *p*-value = 0.837

Thus, we **do not reject** the null hypothesis of equivalent reliabilities

BUT

- Statistical significance is a function of sample size
- These results do not allow for time-varying conclusions
- These tests rely on the proportional hazards assumption

ENTER: Probability of Agreement

PROBABILITY OF AGREEMENT

- The comparison of two groups is often carried out via two-sample hypothesis tests or hypothesis tests that evaluate the need for separate models vs. a single joint model
- Commonly, the null hypothesis associated with such tests assumes that the groups are the same and evidence is sought for dissimilarity
- However, it is often the case that a baseline assumption of inequivalence is more appropriate, in which case evidence is sought for equivalence
- Such is the philosophy of *equivalence testing*[6]

PROBABILITY OF AGREEMENT

- The probability of agreement (PoA) seeks to answer the same question, but without a formal hypothesis test
- Rather, the PoA provides an explicit quantification that the two groups are practically equivalent
- This requires a notion of practical equivalence and an equivalence margin within which differences are considered practically negligible
- This methodology can be broadly applied to a variety of scenarios including
  - The comparison of measurement systems
  - The comparison of fitted or predicted response surfaces
  - The comparison of population reliabilities

PROBABILITY OF AGREEMENT

- For the comparison of KM curves, we define the probability of agreement as

$$\theta(t) = \Pr\left(\left|\hat{S}_1(t) - \hat{S}_2(t)\right| \leq \delta\right)$$

  where $(-\delta, \delta)$ represents an *indifference region* within which differences in reliability are considered practically negligible.

- Thus, it is the probability that the differences between reliabilities, at a given time point t, are practically equivalent.

- Large values of $\theta(t)$ indicate strong agreement while small values signify disagreement.

PROBABILITY OF AGREEMENT

- This probability of agreement, and associated confidence intervals, are visualized as step functions, in a manner similar to KM estimates

- We explore three methods of calculating the probability and associated confidence intervals:
  - Asymptotic normal theory
  - Ordinary bootstrapping
  - Fractional random-weight bootstrapping

THE ASYMPTOTIC APPROACH

Using asymptotic normal theory the probability of agreement becomes a standard normal probability calculation:

For $j = 1,2$ the KM estimator has the following asymptotic distribution:

$$\hat{S}_j(t) \sim N\left(S_j(t), \hat{S}_j(t)^2 \sum_{i:t_{ij}<t} \frac{d_{ij}}{n_{ij}(n_{ij} - d_{ij})}\right)$$

And so:

$$\hat{S}_1(t) - \hat{S}_2(t) \sim N\left(S_0(t), Var\left(\hat{S}_0(t)\right)\right)$$

where: $S_0(t) = S_1(t) - S_2(t)$

$$Var\left(\hat{S}_0(t)\right) = \hat{S}_1(t)^2 \sum_{i:t_{i1}<t} \frac{d_{i1}}{n_{i1}(n_{i1}-d_{i1})} + \hat{S}_2(t)^2 \sum_{i:t_{i2}<t} \frac{d_{i2}}{n_{i2}(n_{i2}-d_{i2})}$$

THE ASYMPTOTIC APPROACH

Then

$$\theta(t) = \Pr\left(\left|\hat{S}_1(t) - \hat{S}_2(t)\right| \leq \delta\right)$$

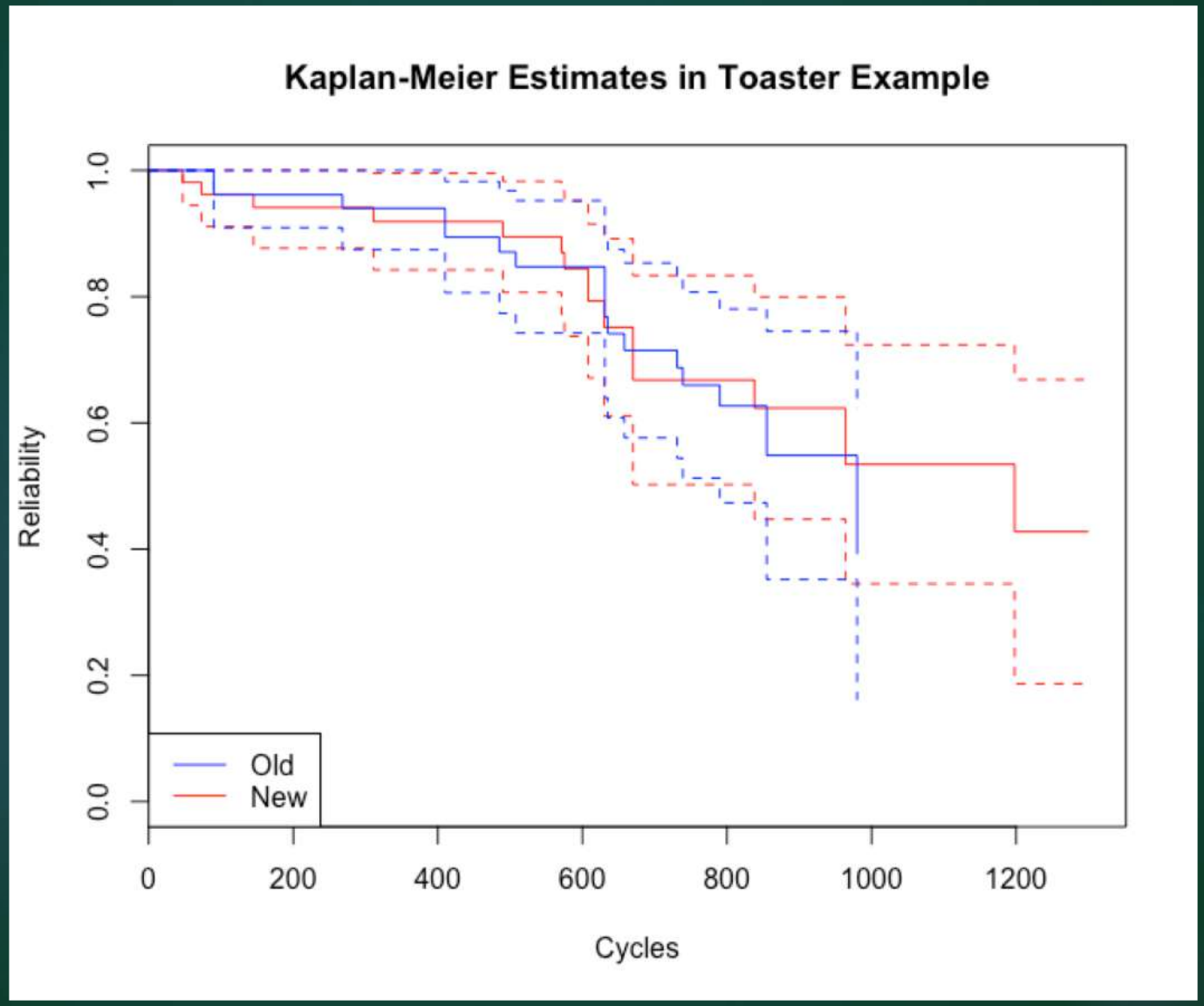
$$= \Phi\left(\frac{\delta - S_0(t)}{\sqrt{Var\left(\hat{S}_0(t)\right)}}\right) - \Phi\left(\frac{-\delta - S_0(t)}{\sqrt{Var\left(\hat{S}_0(t)\right)}}\right)$$

And an approximate $(1-\alpha)\times 100\%$ CI is given by:

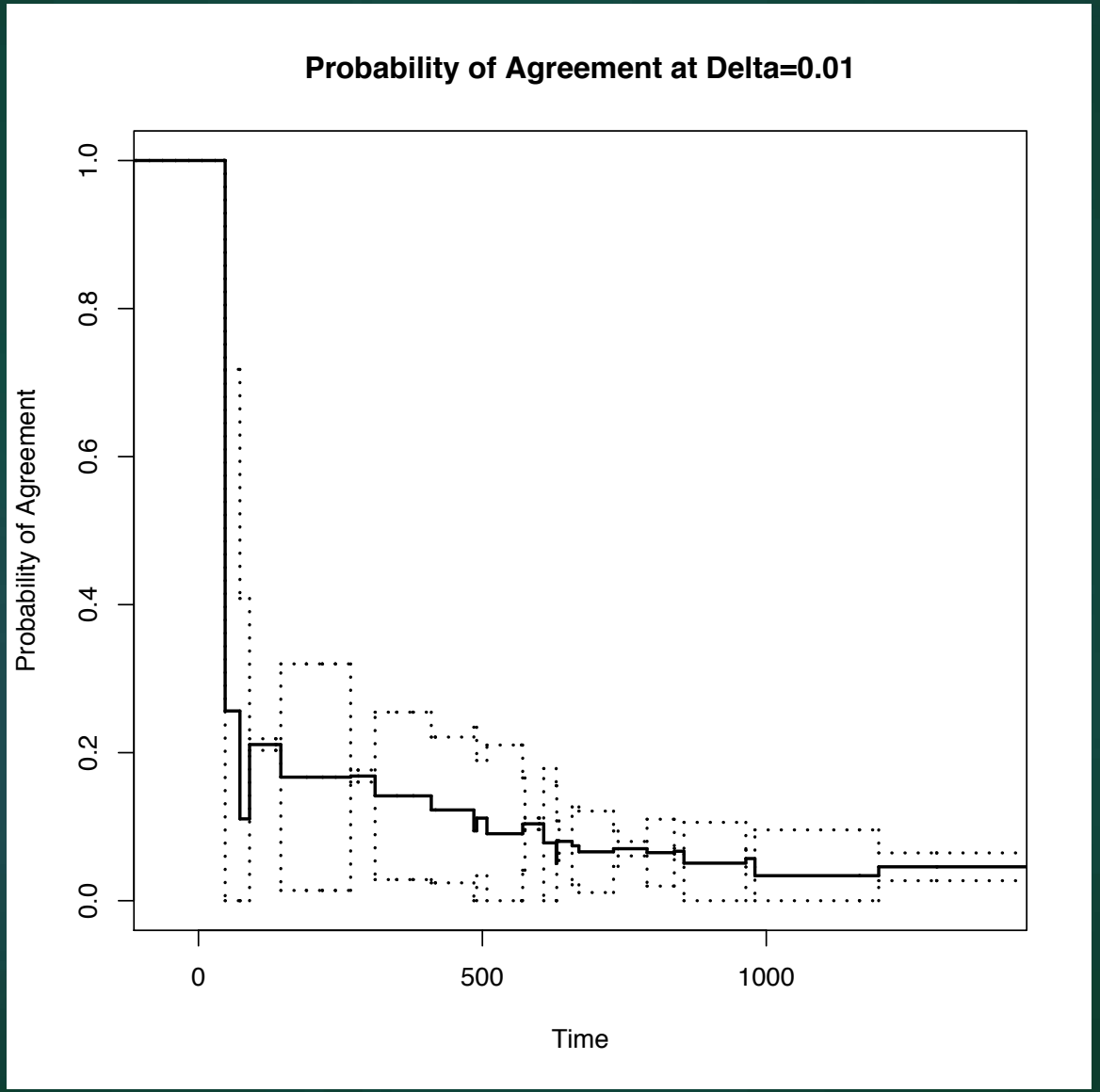
$$\hat{\theta}(t) \pm z_{\alpha/2}\sqrt{Var\left(\hat{\theta}(t)\right)}$$

where $Var\left(\hat{\theta}(t)\right)$ is determined by the Delta Method.

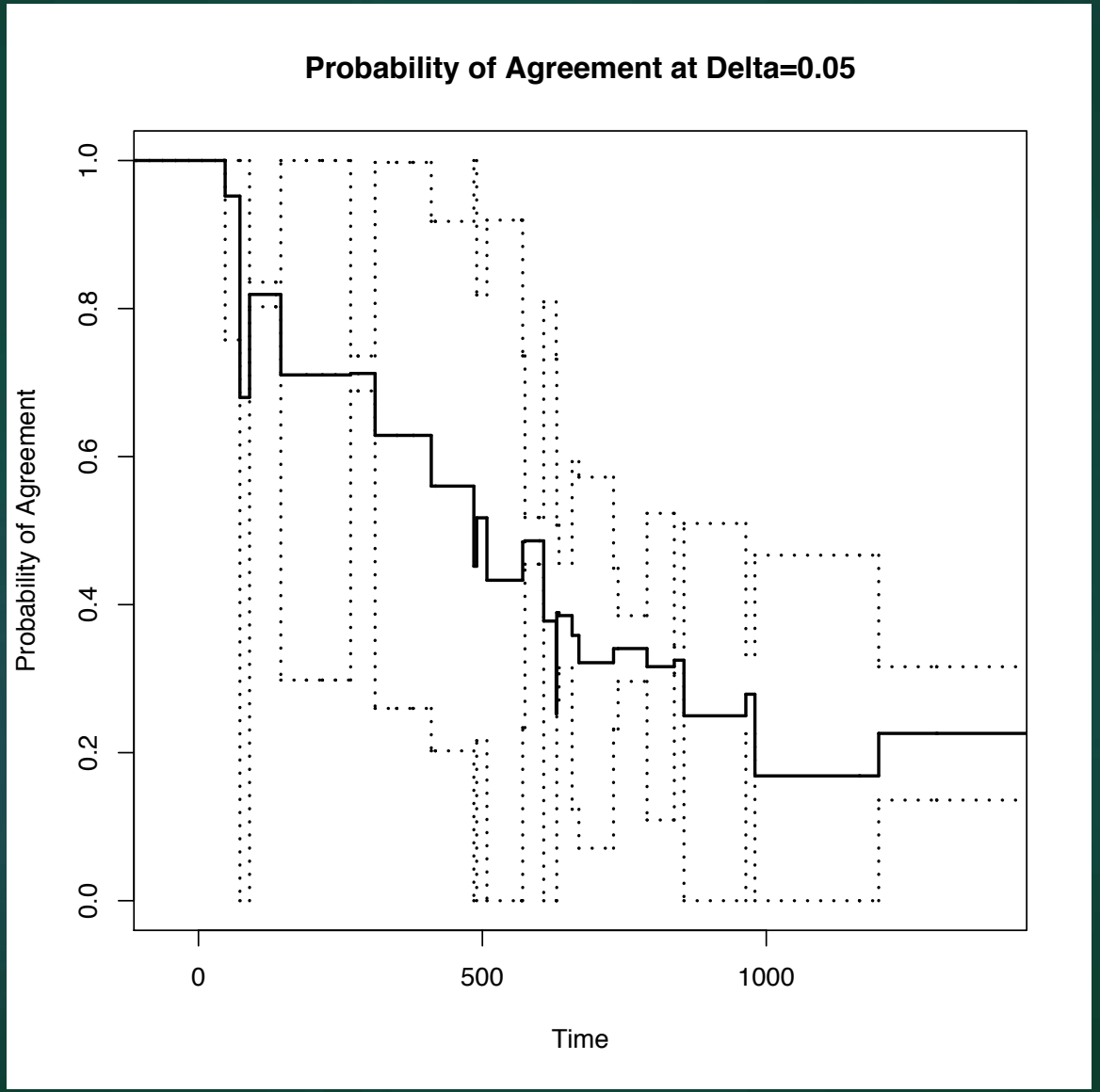
The Toaster Snubber Example



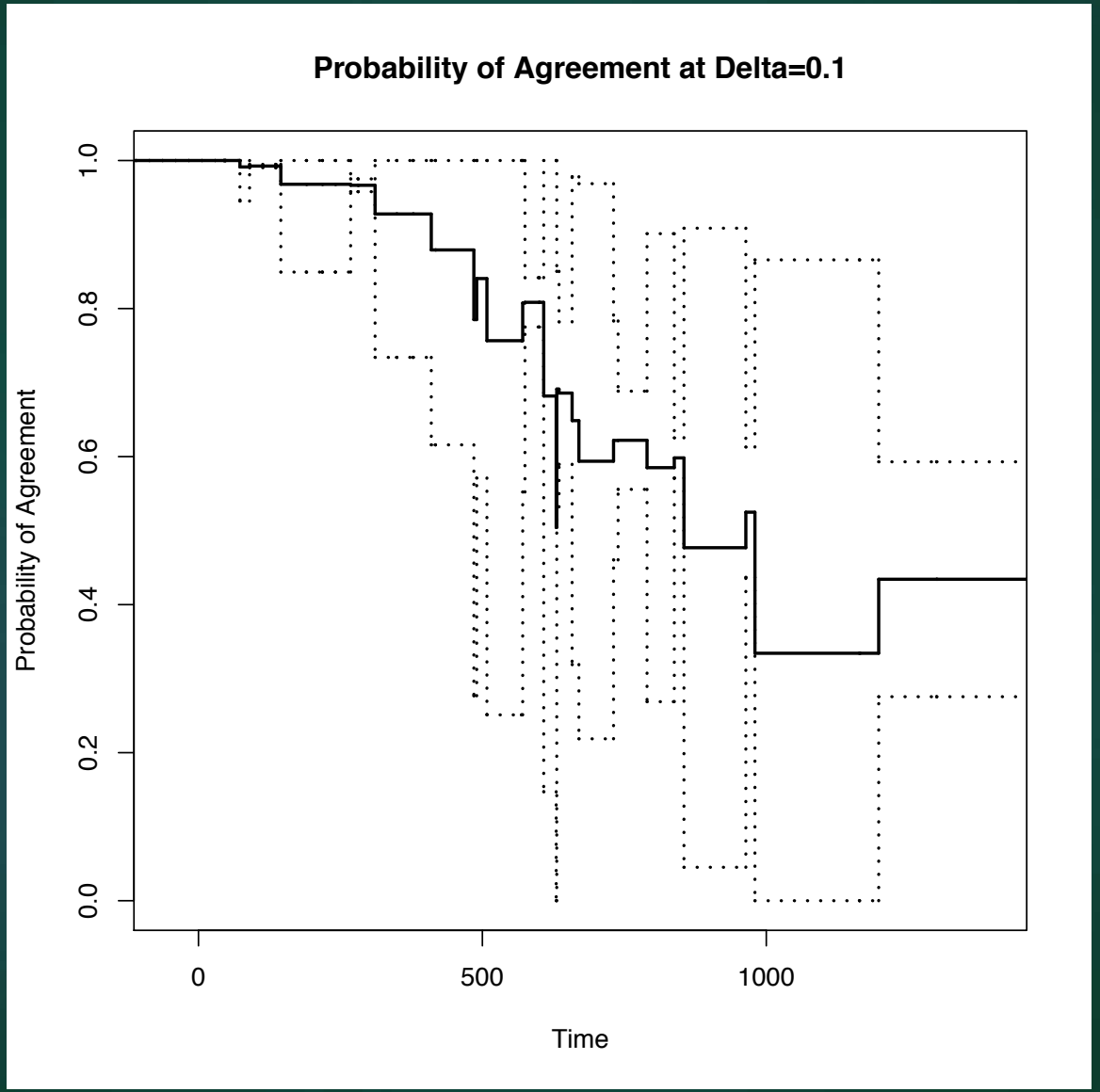
The Toaster Snubber Example



The Toaster Snubber Example



The Toaster Snubber Example



ORDINARY BOOTSTRAPPING

We also consider bootstrapping as a nonparametric alternative that does not rely on large sample sizes or any distributional assumptions

- The Data: pairs of observations (t_{ij}, c_{ij}) where t_{ij} is the event time for unit i in group j and c_{ij} is a censoring indicator ($c_{ij} = 1$ indicates a censored observation; $c_{ij} = 0$ indicates a failure) for $i = 1,2, \dots, r_j$, $j = 1,2$
- What we do in group j: sample pairs with replacement r_j times to obtain a bootstrapped sample (t_{ij}^*, c_{ij}^*), $i = 1,2, \dots, r_j$
- Then we: calculate the KM-estimate $\hat{S}_j^*(t)$ and repeat this process N times, thereby obtaining $\hat{S}_{j1}^*(t), \hat{S}_{j2}^*(t), \dots, \hat{S}_{jN}^*(t)$

ORDINARY BOOTSTRAPPING

Then, for $t \in [0, \max(\{t_{i1}\}, \{t_{i2}\})]$ the probability of agreement is calculated as follows:

$$\hat{\theta}(t) = \frac{1}{N} \sum_{l=1}^{N} I\left\{\left|\hat{S}_{1l}^{*}(t) - \hat{S}_{2l}^{*}(t)\right| \leq \delta\right\}$$

Confidence intervals are calculated by repeating this entire procedure M times so as to obtain M estimates of $\theta(t)$ and hence a sampling distribution for $\hat{\theta}(t)$

Upper and lower confidence bounds are calculated as the $(\alpha/2)\times 100^{\text{th}}$ and $(1-\alpha/2)\times 100^{\text{th}}$ percentiles of this bootstrapped sampling distribution

These point and interval estimates are again plotted as a step function

ORDINARY BOOTSTRAPPING

This approach works well as long as the censoring rate is not too high:

- Because sampling is done with replacement, some of the original observations may not be included in a bootstrapped resample
- This is particularly problematic if the censoring rate is very high and observed failures are very rare
- In this case, it is possible that a bootstrapped resample may contain too few observed failures, leading to an inaccurate estimate of $S_j(t)$
- To avoid this problem we consider fractional random-weight bootstrapping[7]

FRACTIONAL RANDOM-WEIGHT BOOTSTRAP

Sampling with replacement in the ordinary bootstrap case is equivalent to randomly assigning non-negative integer weights to each observation (t_{ij}, c_{ij}), $i = 1,2, \ldots, r_j$ where the weights sum to r_j.

Thus a weight of zero means that the corresponding observation is not included in the resample.

To avoid the problems described previously, we could instead assign positive fractional random weights (FRW) that also sum to r_j but that, by definition, cannot be zero

We illustrate this idea by considering the Toaster Snubber Example

FRACTIONAL RANDOM-WEIGHT BOOTSTRAP

Obs.	t_{i1}	c_{i1}	OW	FRW
1	90	0	0	2.1109
2	90	0	2	2.0084
3	90	1	2	0.3900
4	190	1	0	0.7179
5	218	1	0	2.3111
6	218	1	2	0.5217
7	241	1	1	0.6698
8	268	0	4	1.4360
⋮	⋮	⋮	⋮	⋮
48	980	1	1	0.8423
49	980	1	0	0.3940
50	980	1	3	3.6417
51	980	1	1	0.1916
52	980	1	1	1.8151

FRACTIONAL RANDOM-WEIGHT BOOTSTRAP

t_{i1}	n_{i1}	d_{i1}	$\hat{S}_1(t)$
0	52	0	1
90	52	2.1109 + 2.0084 = 4.1193	0.9208
268	52 – 8.7298 = 43.2702	1.4360	0.8902
410	43.2702 – 3.2069 = 40.0633	1.3271 + 0.4307 = 1.7578	0.8511
485	40.0633 – 1.9708 = 38.0925	0.9610	0.8296
508	38.0925 – 0.9610 = 37.1315	0.4494	0.8196
631	37.1315 – 5.4507 = 31.6808	0.4591 + 1.3811 + 0.2808 = 2.1210	0.7647
635	31.6808 – 2.1210 = 29.5598	1.5626	0.7243
⋮	⋮	⋮	⋮

FRACTIONAL RANDOM-WEIGHT BOOTSTRAP

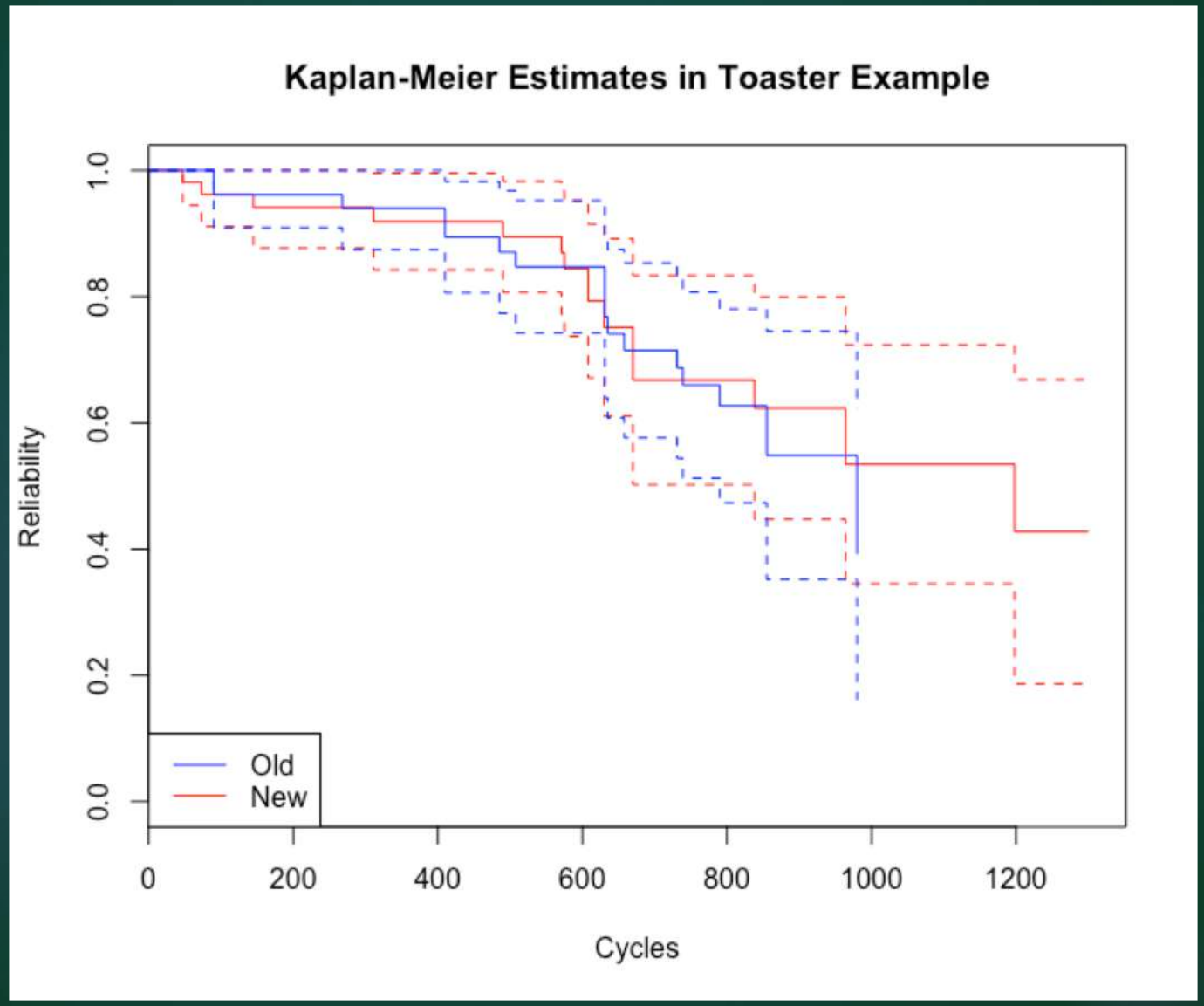
A bootstrap sample with this approach is taken by weighting each observation (failure or censored) according to a random assignment of FRWs such as those shown in the previous table.

In order to calculate the probability of agreement we obtain N such fractional random-weight bootstrap samples for each group and with them obtain N KM estimates of $S_1(t)$ and $S_2(t)$.

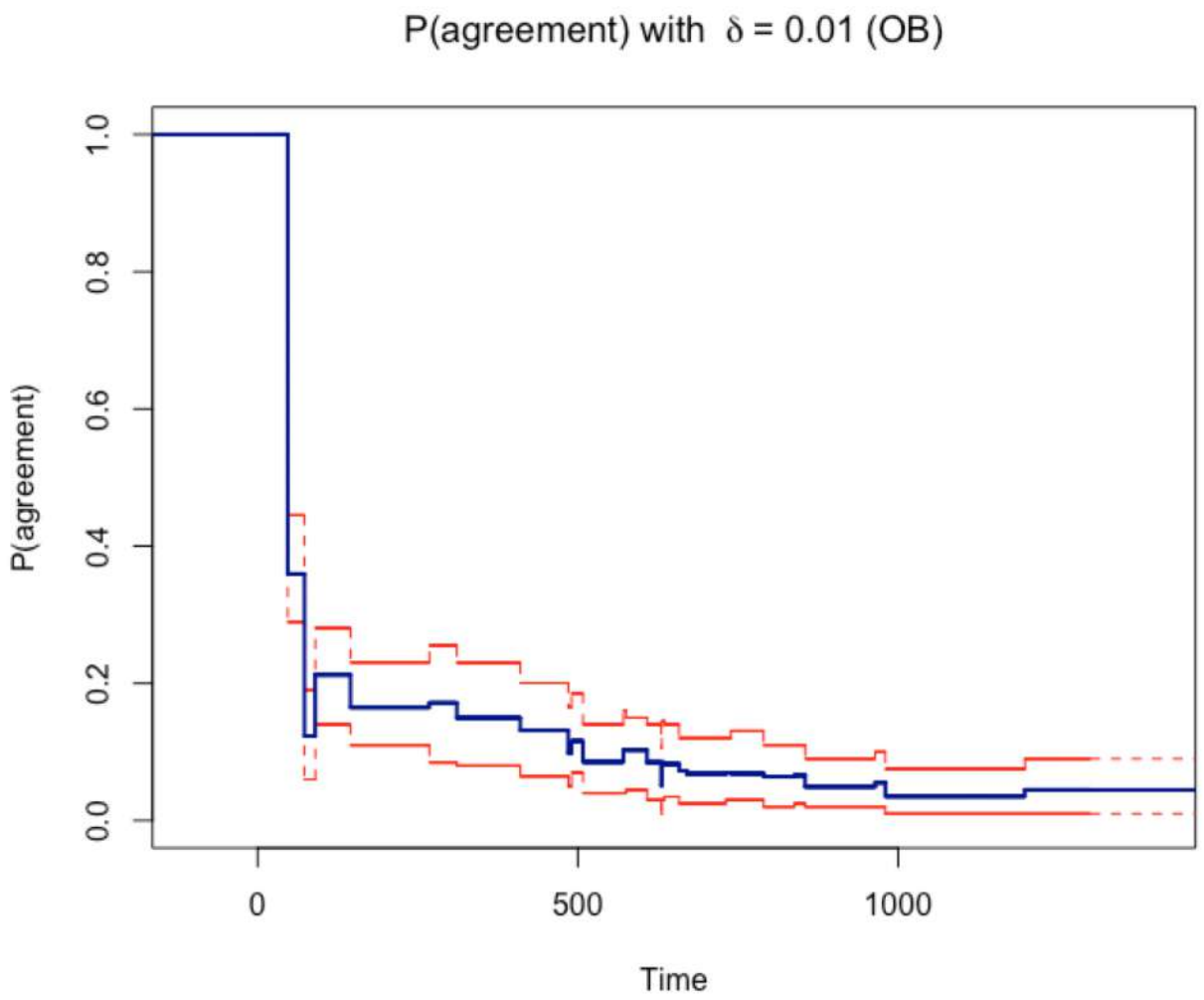
The estimate $\hat{\theta}(t)$ and associated confidence intervals are both calculated as in the ordinary bootstrap case:

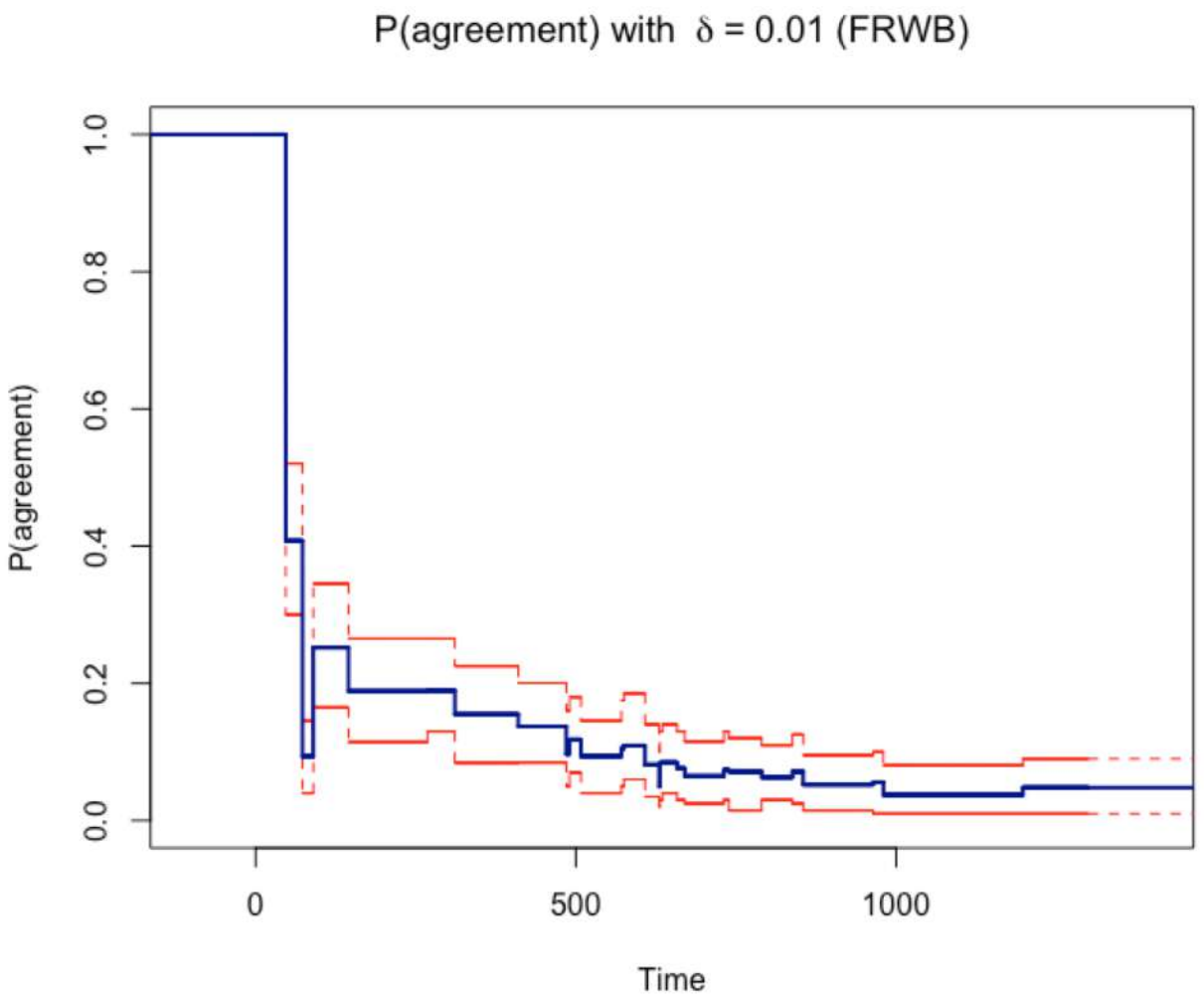
$$\hat{\theta}(t) = \frac{1}{N}\sum_{l=1}^{N} I\left\{\left|\hat{S}_{1l}^{*}(t) - \hat{S}_{2l}^{*}(t)\right| \leq \delta\right\}$$

The Toaster Snubber Example

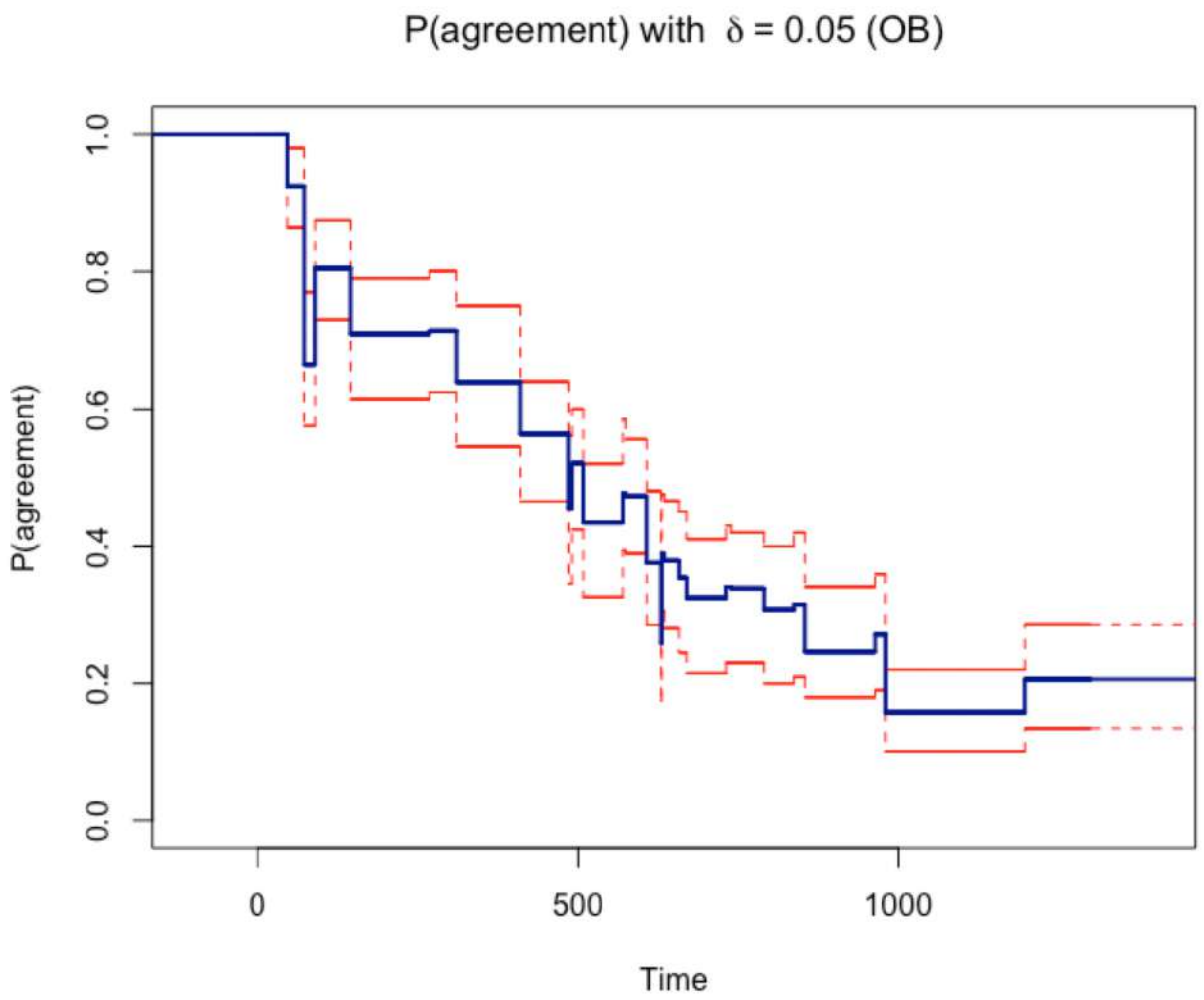


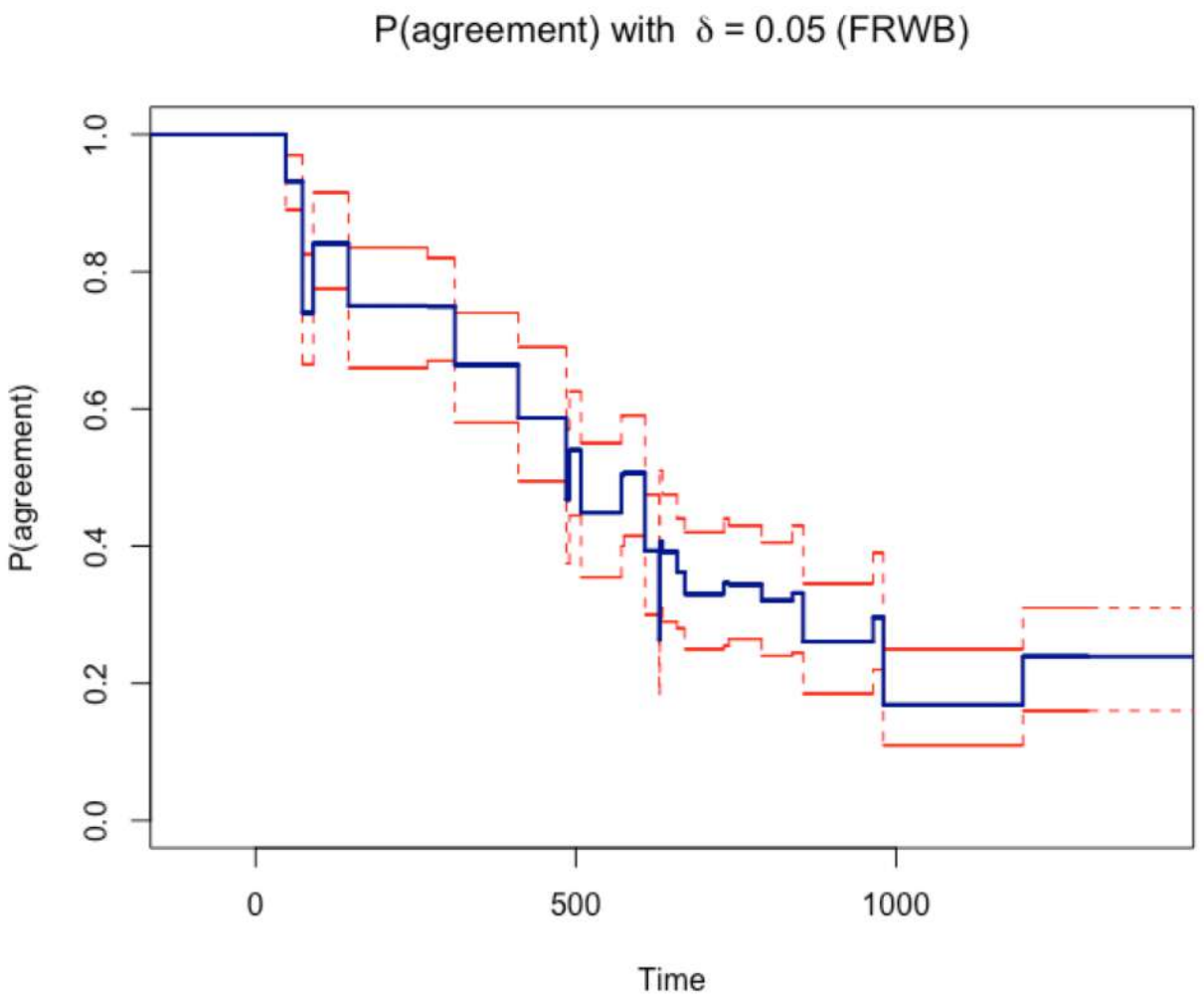
The Toaster Snubber Example



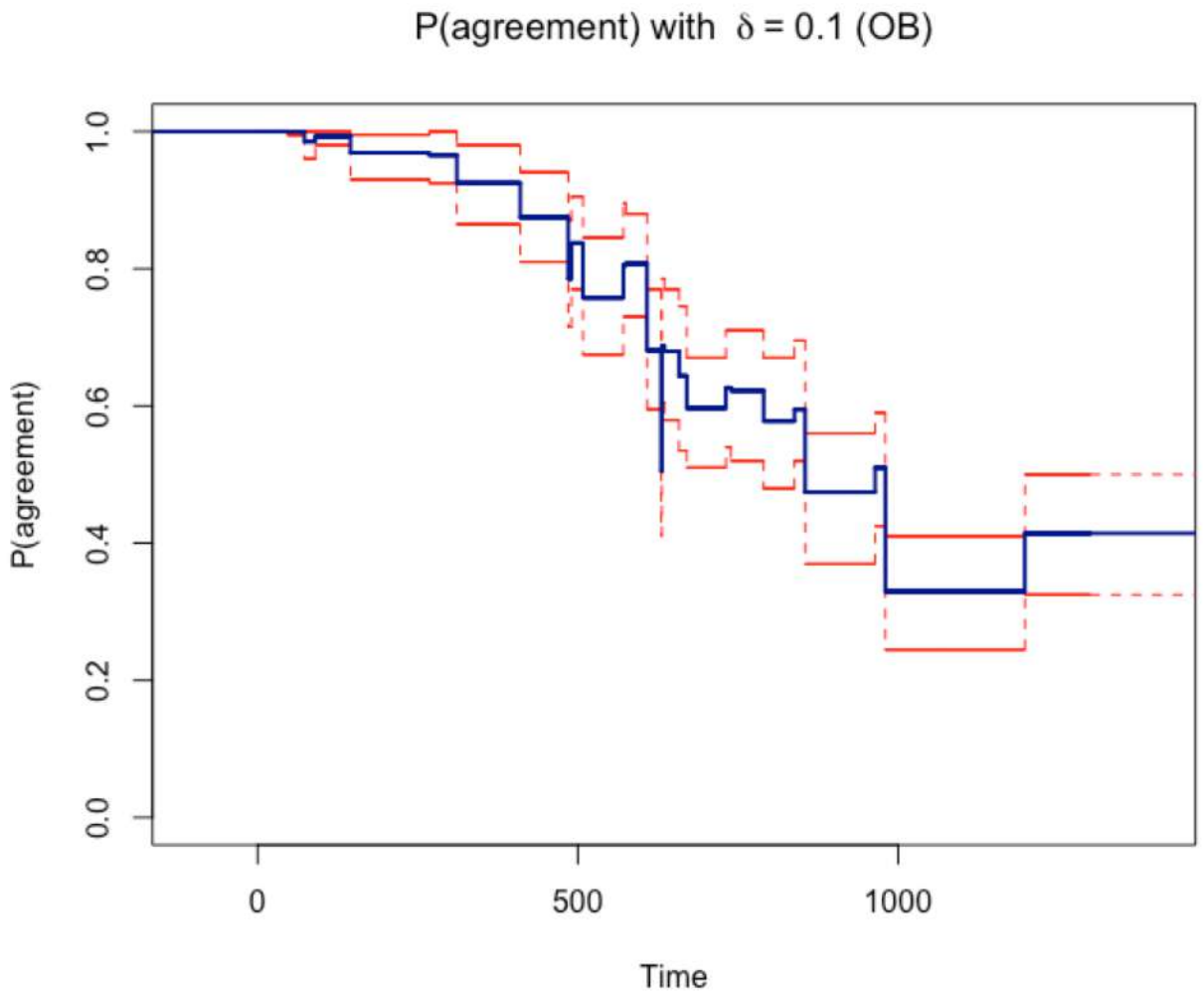


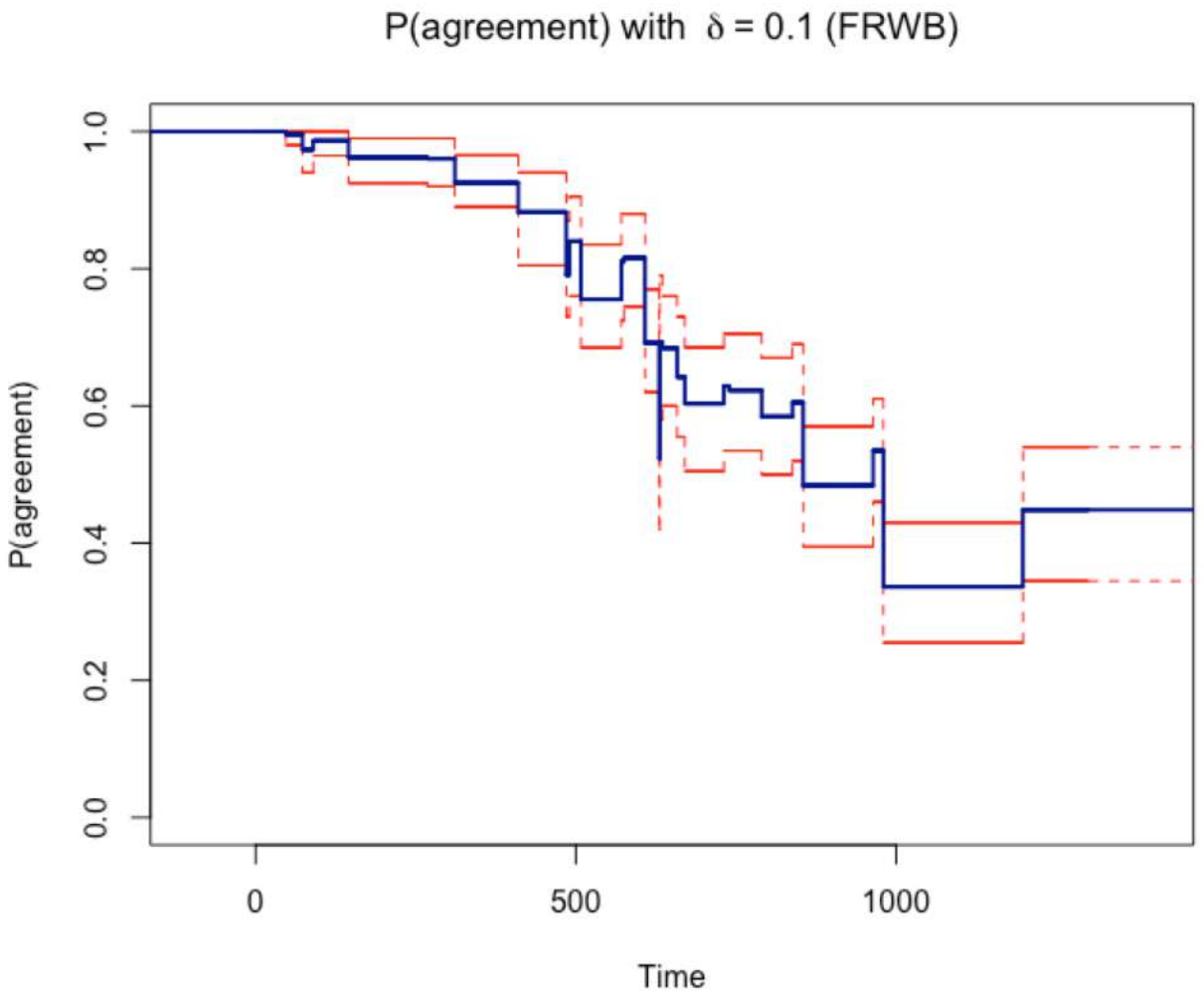
The Toaster Snubber Example





The Toaster Snubber Example





The Toaster Snubber Example

What did we find?

1. As the number of cycles increases, agreement steadily decreases
2. The timing and magnitude of this disagreement depends on δ
3. Agreement increases slightly for very large numbers of cycles
4. Confidence intervals are narrower for bootstrap approaches
5. Asymptotic and bootstrap estimates of $\theta(t)$ are similar when the number of events $(r_1 + r_2)$ is large
6. Ordinary and FRW bootstrapping give similar results as long as the censoring rate is not too high

SUMMARY

The probability of agreement provides an intuitive and practically useful means of comparing reliabilities in two populations

- It facilitates time-dependent conclusions
- Conclusions are not predetermined by sample size
- The proportional hazards assumption does not need to be made

Whether one decides that the reliability in two populations is sufficiently similar to combine them, and use a single reliability model, requires practical decisions made by the practitioner:

- How different is too different?
- How large a value of the PoA is large enough?

THANK YOU!

REFERENCES

1. Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282): 457-481.
2. Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22(4): 719-748.
3. Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52: 203-224.
4. Harrington, D.P. and Fleming, T.R. (1982). A class of rank test procedures for censored survival data. *Biometrika* 69(3): 553-566.
5. Nelson, W.B. (1984). *Accelerated testing: statistical models, test plans, and data analysis*. John Wiley & Sons, Inc.
6. Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*, 2nd ed. Chapman and Hall/CRC Press.
7. Meeker, W.Q., Hahn, G.J., and Escobar, L.A. (2017) Statistical Intervals: A Guide for Practitioners and Researchers, John Wiley & Sons, Inc.