Bayesian probability of agreement for comparing survival or reliability functions with parametric lifetime regression models

NATHANIEL T. STEVENS

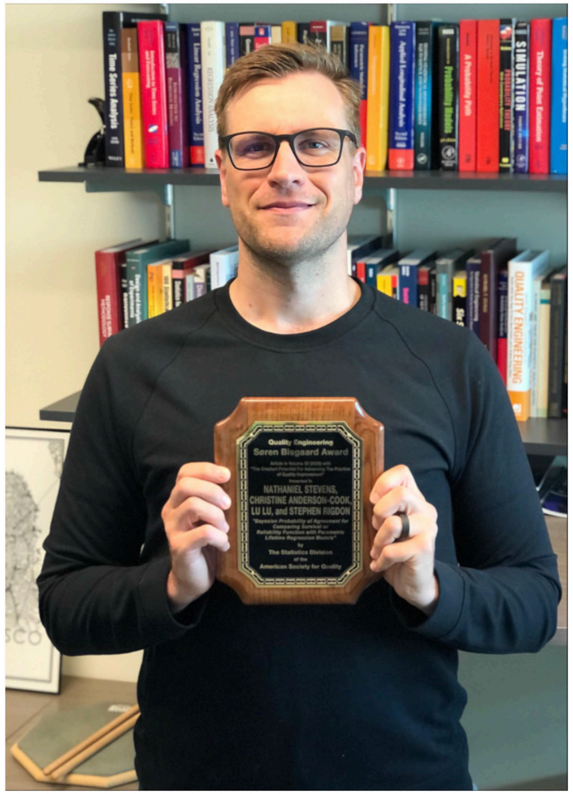
FALL TECHNICAL CONFERENCE WEBINAR SERIES

OCTOBER 15, 2021

UNIVERSITY OF WATERLOO

Acknowledgements

- Stevens, N. T., Lu, L., Anderson-Cook, C. M., & Rigdon, S. E. (2020). Bayesian probability of agreement for comparing survival or reliability functions with parametric lifetime regression models. *Quality Engineering*, *32*(3), 312-332.



Acknowledgements

Søren Bisgaard (1951-2009)

- Bachelor's (1975) and Master's (1979) degrees in engineering and PhD (1985) in statistics.
- Prolific author and researcher in industrial statistics and quality engineering, who emphasized practical problem solving.
- His impact and importance was recognized by several awards including the Wilcoxon Prize, Shewell Award, Brumbaugh Award, Shewhart Medal, George Box Medal, ASQ and ASA Fellowships.
- Instrumental in establishing ENBIS.

Outline

- Problem Description
- Practical Equivalence
- The (Bayesian) Probability of Agreement
- Examples
- Summary

Problem Description

Problem Description

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

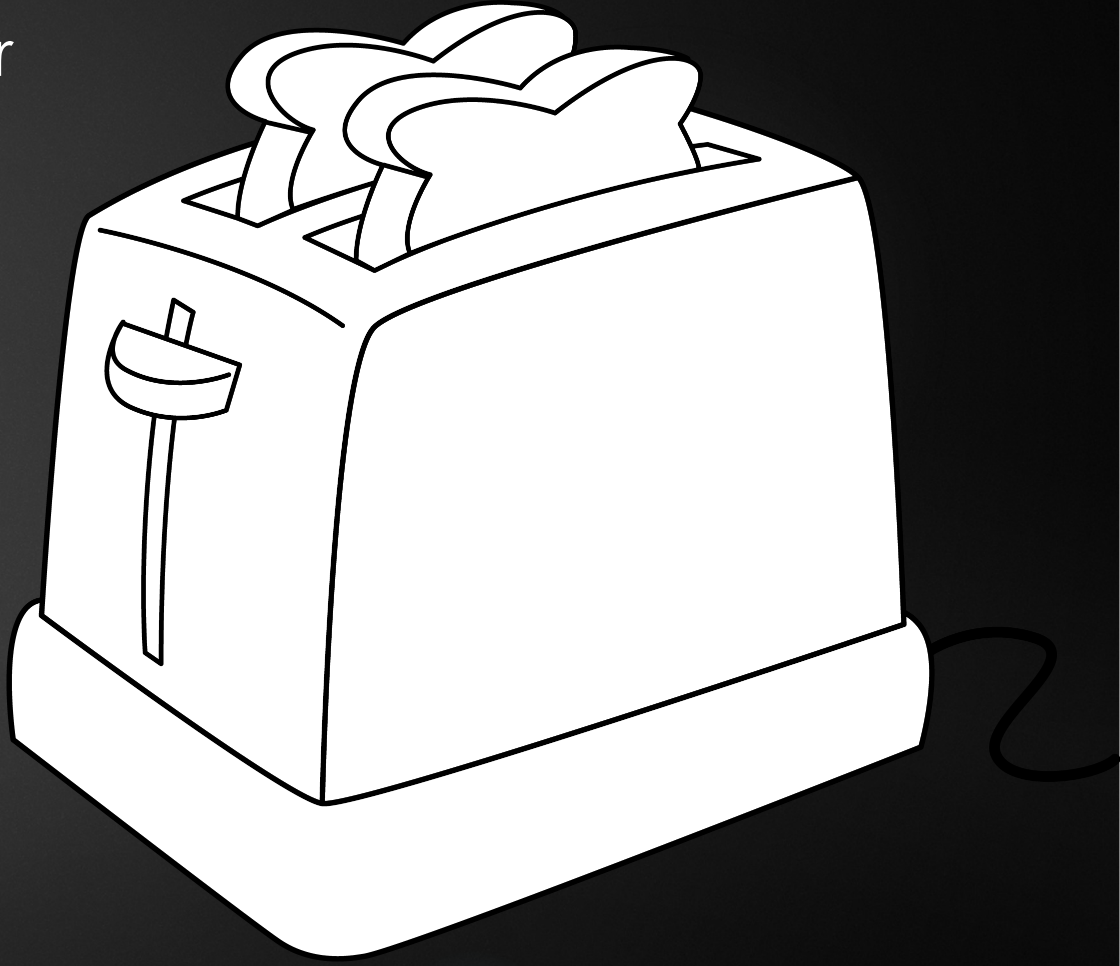
Contextual goal: compare reliabilities of two related populations

Problem Description

Nelson[1] 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 $n_1 = 52$ “old” toasters and $n_2 = 54$ “new” toasters involved in this comparison.

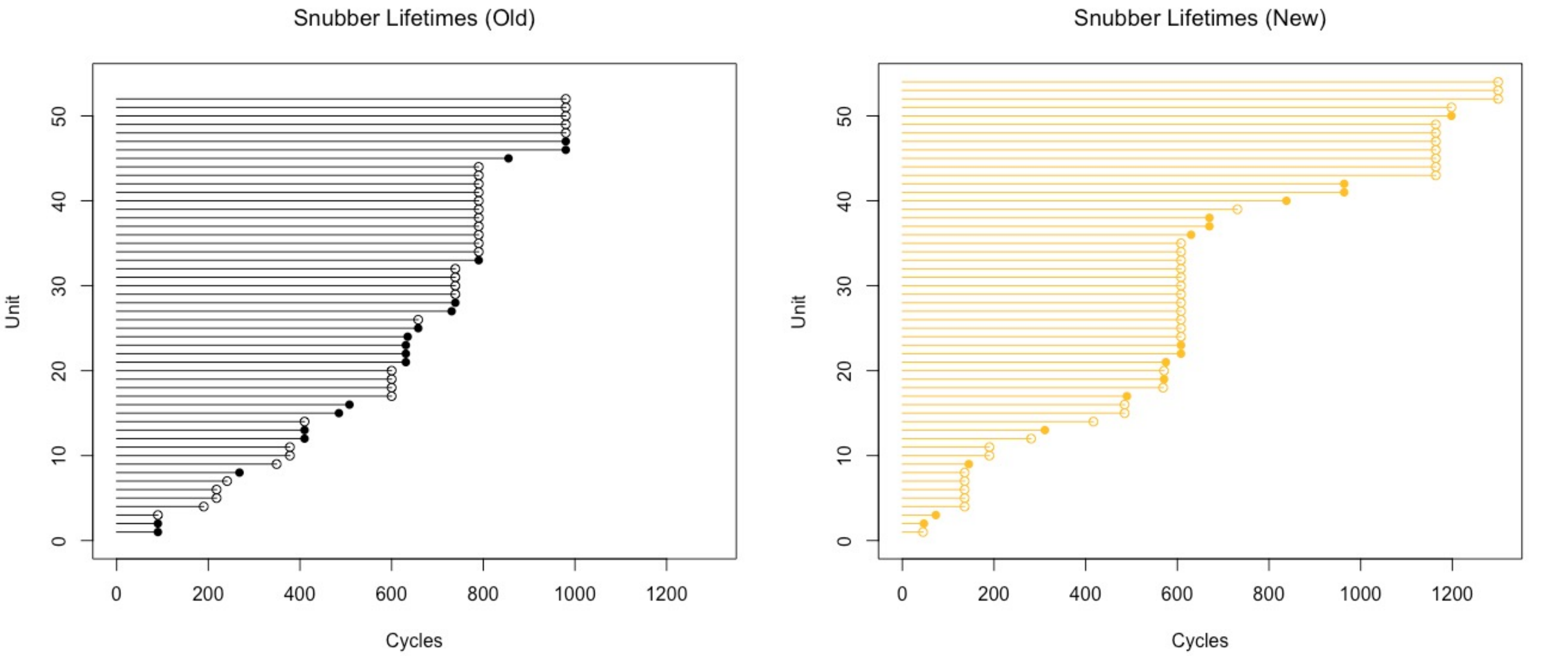


Problem Description

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+	

Problem Description



Practical Equivalence

Practical Equivalence

- 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*[2]

Practical Equivalence

- Central to this philosophy is the understanding that two quantities don't need to be *identical* for them to be *practically equivalent*.
- This notion acknowledges that their exists a size of difference that is practically unimportant.
- Methodologies that emphasize practical importance over statistical significance are gaining popularity in the wake of the current p-value controversy and reproducibility crisis[3].
- Here we propose a methodology for the comparison of parametric lifetime regressions that explicitly accounts for the notion of practical equivalence, and that is rooted in Bayesian estimation.

Probability of Agreement

Probability of Agreement

- We extend the use of the probability of agreement (PA) to the comparison of lifetime distributions with censored data.

- In general, given characteristics θ_1 and θ_2, the PA explicitly quantifies the likelihood that θ_1 and θ_2 are practically equivalent:

$$PA = \Pr(|\theta_1 - \theta_2| < \delta)$$

  where $\delta > 0$ is the equivalence margin and $(-\delta, \delta)$ is the region of practical equivalence, within which differences are considered practically negligible

- Large PA values indicate strong agreement while small values signify disagreement.

Probability of Agreement

- This methodology has been broadly applied to a variety of scenarios including
  - The comparison of measurement systems[4,5]
  - The comparison of fitted or predicted response surfaces[6,7]
  - The comparison of sequential experimental results[8]

- Here we take θ_1 and θ_2 to be quantities that summarize two lifetime distributions and we use the PA to quantify the likelihood that they are practically equivalent.

- Because θ_1 and θ_2 are parameters (or functions of parameters), the Bayesian paradigm is most appropriate, allowing for intuitive interpretation

Probability of Agreement

- We define the Bayesian probability of agreement (BPA) in this setting as

$$BPA = \Pr(|\theta_1 - \theta_2| < \delta | \text{data})$$

  which is a *posterior* probability calculated given observed data and assuming $(-\delta, \delta)$ is the region of practical equivalence defined earlier.

- Here we assume that $T_{ij} \sim F_j$ is a random variable representing the lifetime of unit i in group j, $i = 1,2, \ldots, n_j$, $j = 1,2$.

- In this work we assume the lifetime distribution F_j is Weibull, lognormal, or gamma, though other distributional assumptions may easily be accommodated.

Probability of Agreement

- We further assume that θ_j is a parameter or function of parameters that usefully describes the lifetime distribution F_j, such as:

$$\theta_j = \Pr\left(T_{ij} \geq t\right) = 1 - F_j(t) \qquad \text{or} \qquad \theta_j = F_j^{-1}(p)$$

- As we can see, θ_1 and θ_2 may themselves be functions of input(s) such as t, p, or other context-dependent covariates $\boldsymbol{x}$.

- Generally speaking, interest lies in comparing $\theta_1 = h(\boldsymbol{x}_1^T\boldsymbol{\beta}_1)$ with $\theta_2 = h(\boldsymbol{x}_2^T\boldsymbol{\beta}_2)$.

- The BPA can then be calculated and visualized across a range of relevant values of the inputs, thereby quantifying the similarity of θ_1 and θ_2 in regions of interest.

Probability of Agreement

- The BPA is straightforward to interpret: it quantifies the strength of evidence in favour of the statement $|\theta_1 - \theta_2| < \delta$
  - Values close to 1 provide strong evidence in favour of this statement
  - Values close to 0 provide strong evidence in favour of this statement's complement

- How large the BPA needs to be in order to believe $|\theta_1 - \theta_2| < \delta$ is determined by the user.

- To ensure practically useful conclusions, δ should be chosen carefully to provide a meaningful comparison in the context of the problem.

Probability of Agreement

- Given the observed data $\left(t_{ij}, c_{ij}, \boldsymbol{x}_{ij}\right)$, $i = 1,2,\ldots,n_j$, $j = 1,2$ and the joint posterior $p(\theta_1, \theta_2 | \boldsymbol{t}, \boldsymbol{c}, \boldsymbol{x})$ the BPA may be calculated as

$$BPA = \iint_{\mathcal{D}} p(\theta_1, \theta_2 | \boldsymbol{t}, \boldsymbol{c}, \boldsymbol{x}) d\theta_1 d\theta_2$$

  where $\mathcal{D}$ is the region for which $|\theta_1 - \theta_2| < \delta$. However, in general, this integral cannot be evaluated analytically.

- For ample flexibility, we approximate $p(\theta_1, \theta_2 | \boldsymbol{t}, \boldsymbol{c}, \boldsymbol{x})$ via MCMC simulation and estimate the BPA as follows:

$$\widehat{BPA} = \frac{1}{M} \sum_{k=1}^{M} \mathbb{I}\{|\theta_{1k} - \theta_{2k}| < \delta\}$$

Probability of Agreement

- The posterior draws $\theta_{j1}, \theta_{j2}, \dots, \theta_{jM}$ (for both $j = 1,2$) are obtained by taking draws from the posteriors of $\boldsymbol{\beta}_{j1}, \boldsymbol{\beta}_{j2}, \dots, \boldsymbol{\beta}_{jM}$ and calculating $\theta_{jk} = h\left(\boldsymbol{x}_j^T \boldsymbol{\beta}_{jk}\right)$ for each $j = 1,2$ and $k = 1,2, \dots, M$.

- Note that the value M is the number of posterior draws retained after a sufficient burn-in and thinning.

- We assume diffuse priors for $\boldsymbol{\beta}_j$ (i.e., $\beta_{0j}, \beta_{1j}, \dots, \beta_{pj} \sim \mathrm{N}(0{,}1000)$) to reflect the assumption that a practitioner may not have strong prior knowledge.

- We use simulation to investigate the effect of the choice of prior

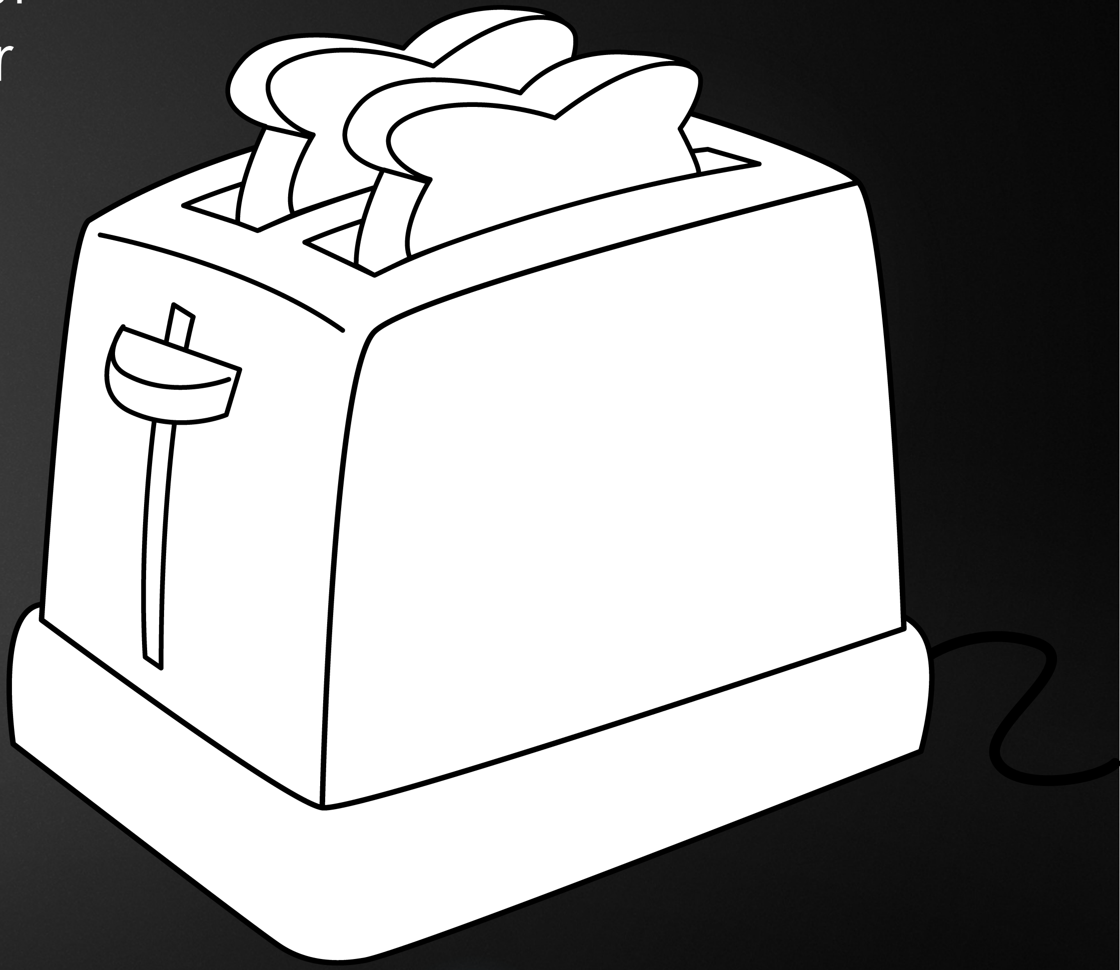
Examples

Toaster Snubber Example

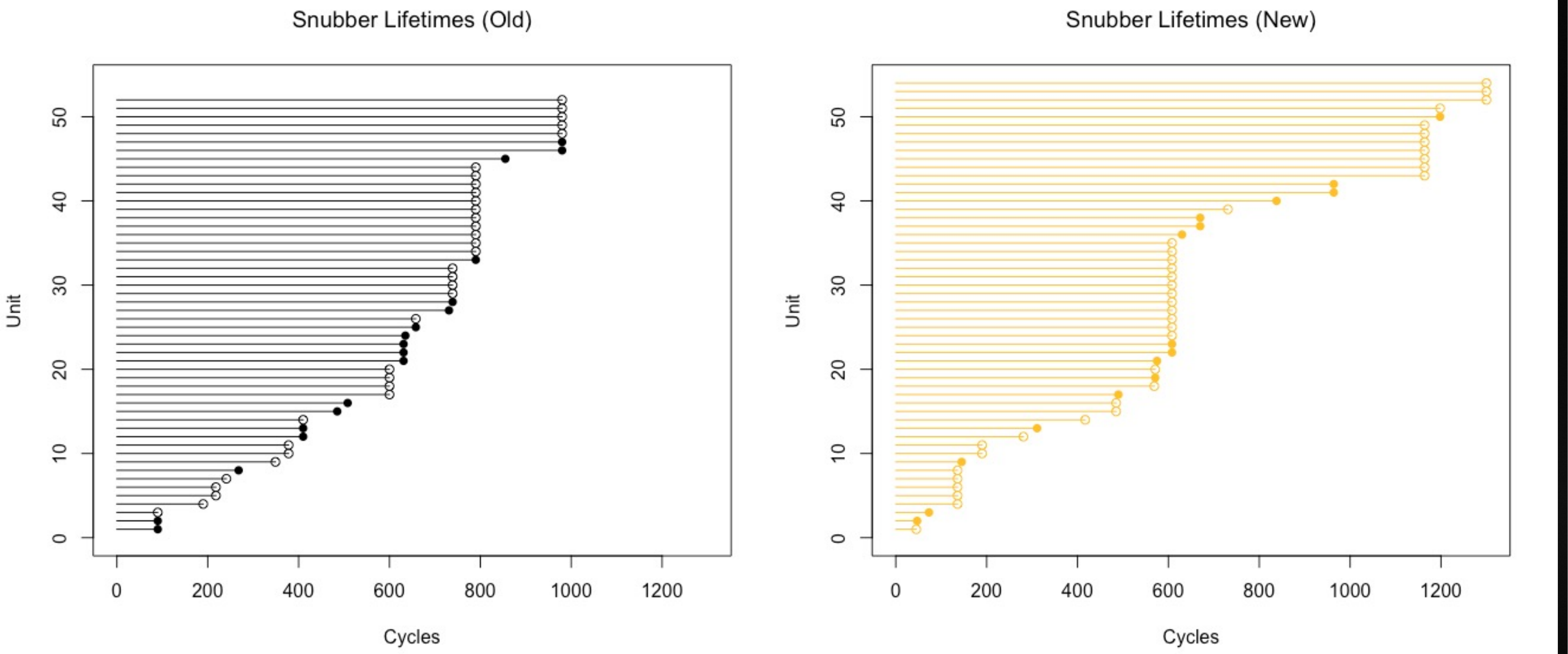
Nelson[1] describes an accelerated life test in which two different versions of a toaster are repeatedly cycled.

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

There were $n_1 = 52$ "old" toasters and $n_2 = 54$ "new" toasters involved in this comparison.

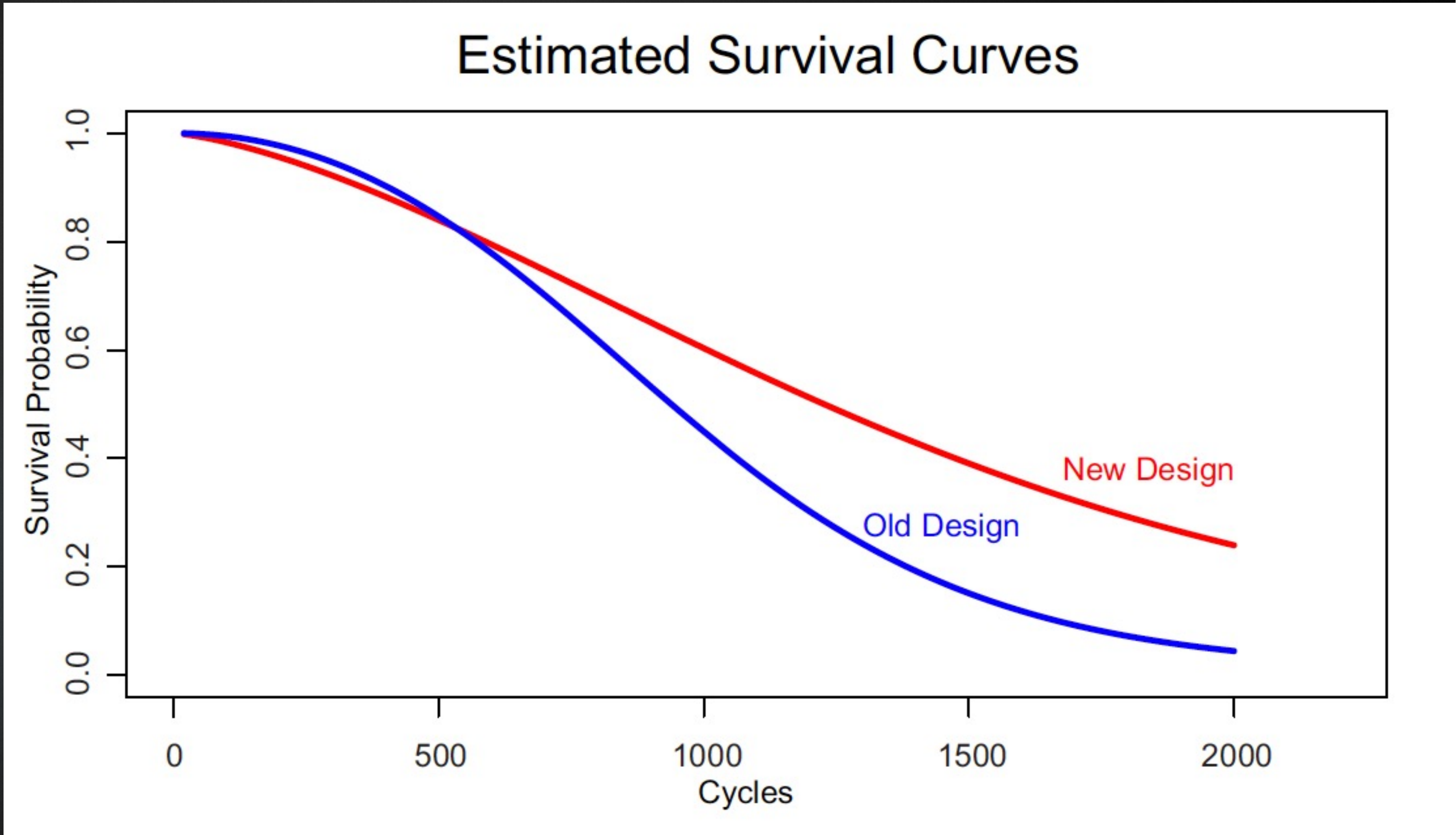


Toaster Snubber Example



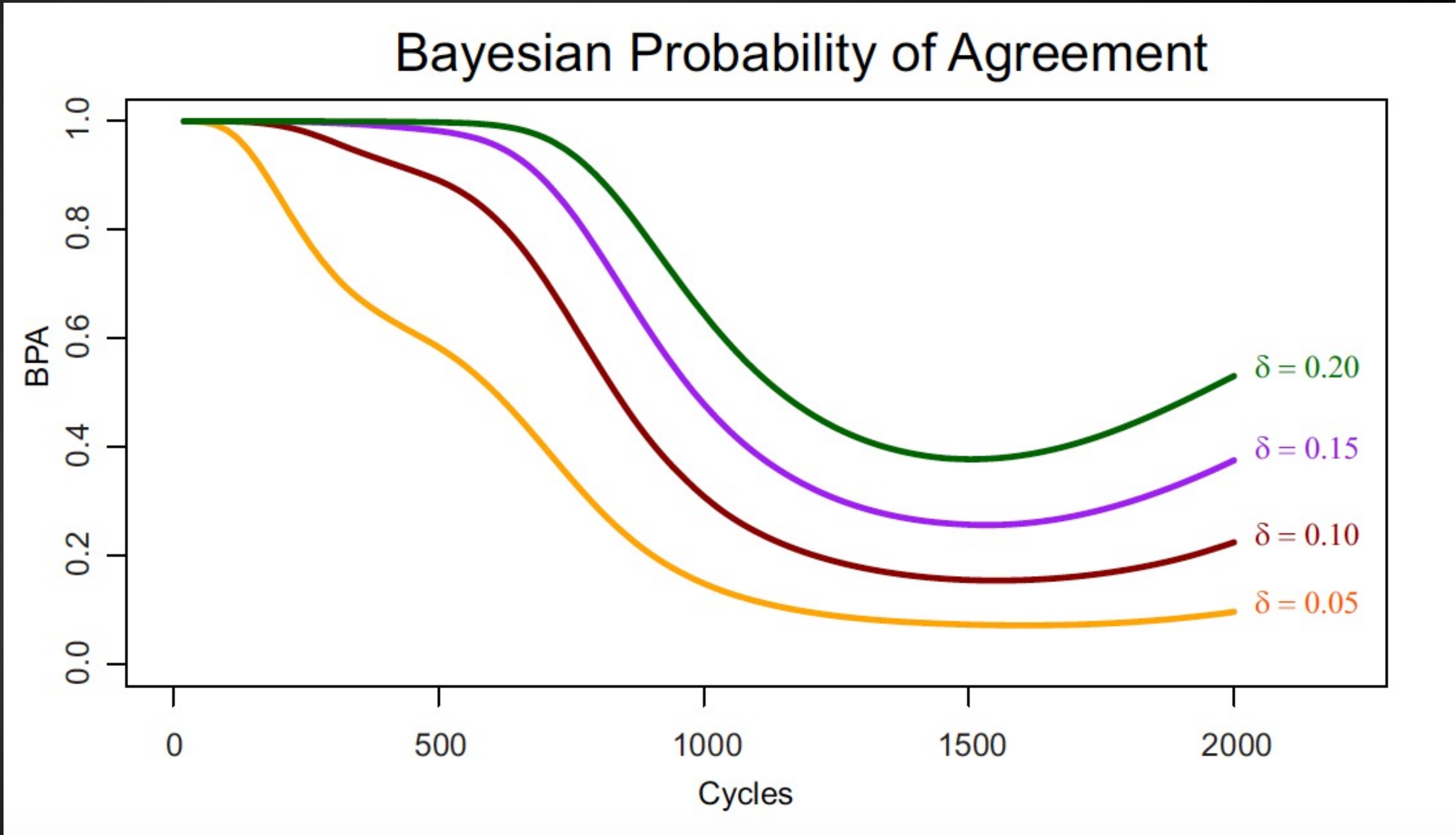
Toaster Snubber Example

$$\theta_j = 1 - F_j(t)$$



Toaster Snubber Example

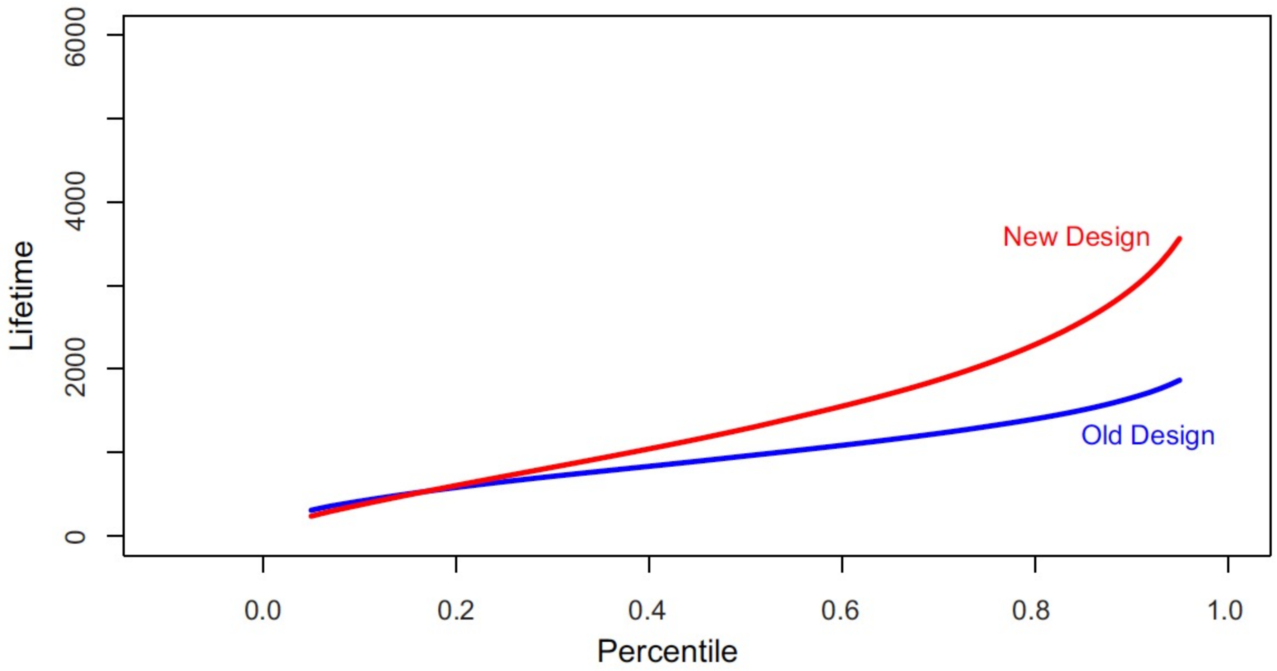
$$\Pr(|\theta_1 - \theta_2| < \delta|\text{data})$$

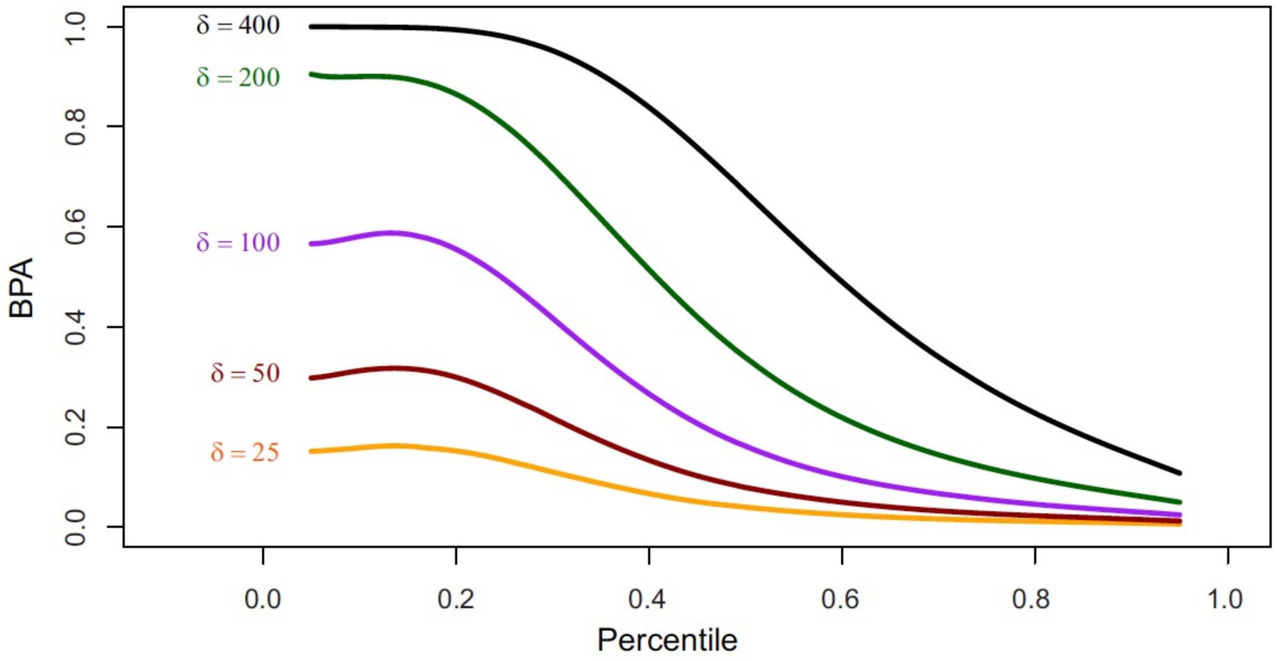


Toaster Snubber

$$\theta_j = F_j^{-1}(p)$$

$$\Pr(|\theta_1 - \theta_2| < \delta|\text{data})$$





Toaster Snubber Example

What did we find?

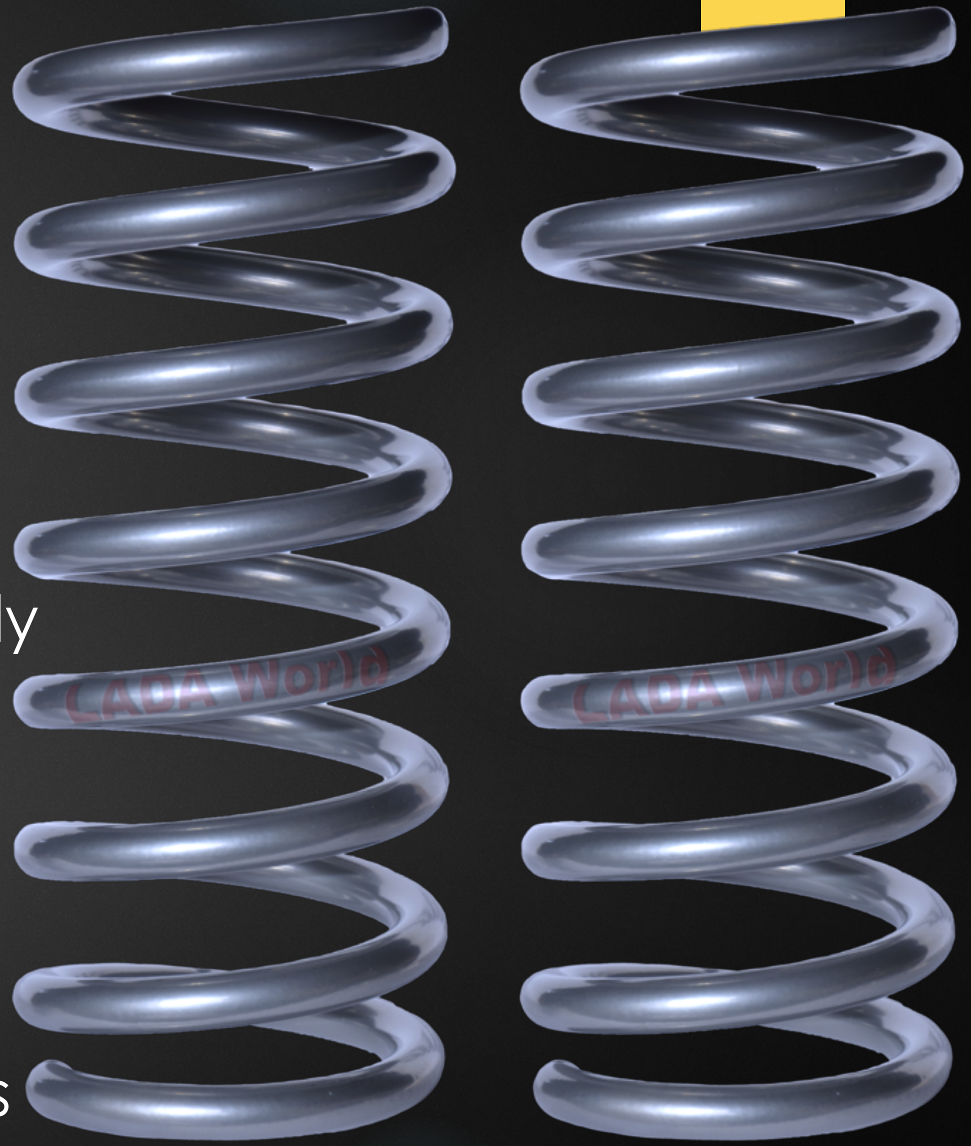
1. As the number of cycles increases, agreement steadily decreases
2. The timing and magnitude of this disagreement depends on δ
3. Agreement increases slightly for very large numbers of cycles

Spring Example

Meeker et al.[9] describes an accelerated life test performed to assess the reliability of a spring under new and old processing methods.

The goal was to determine whether the new processing method would meaningfully improve the spring’s *fatigue life* (the number of kilocycles sustained before failure).

108 ($n_1 = 52$ “old” and $n_2 = 54$ “new”) springs were tested for up to 5000 kilocycles in a 2×2×3 factorial experiment.



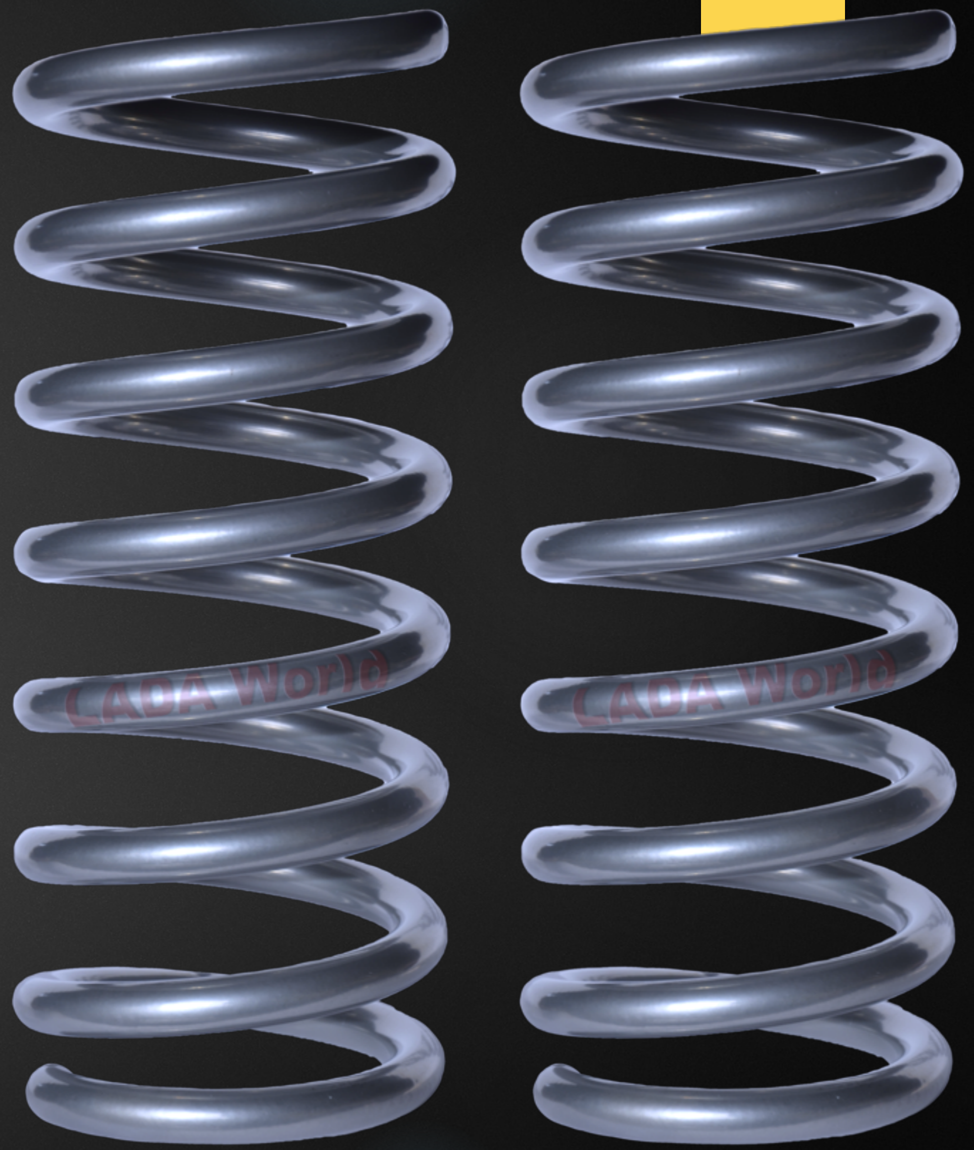
Spring Example

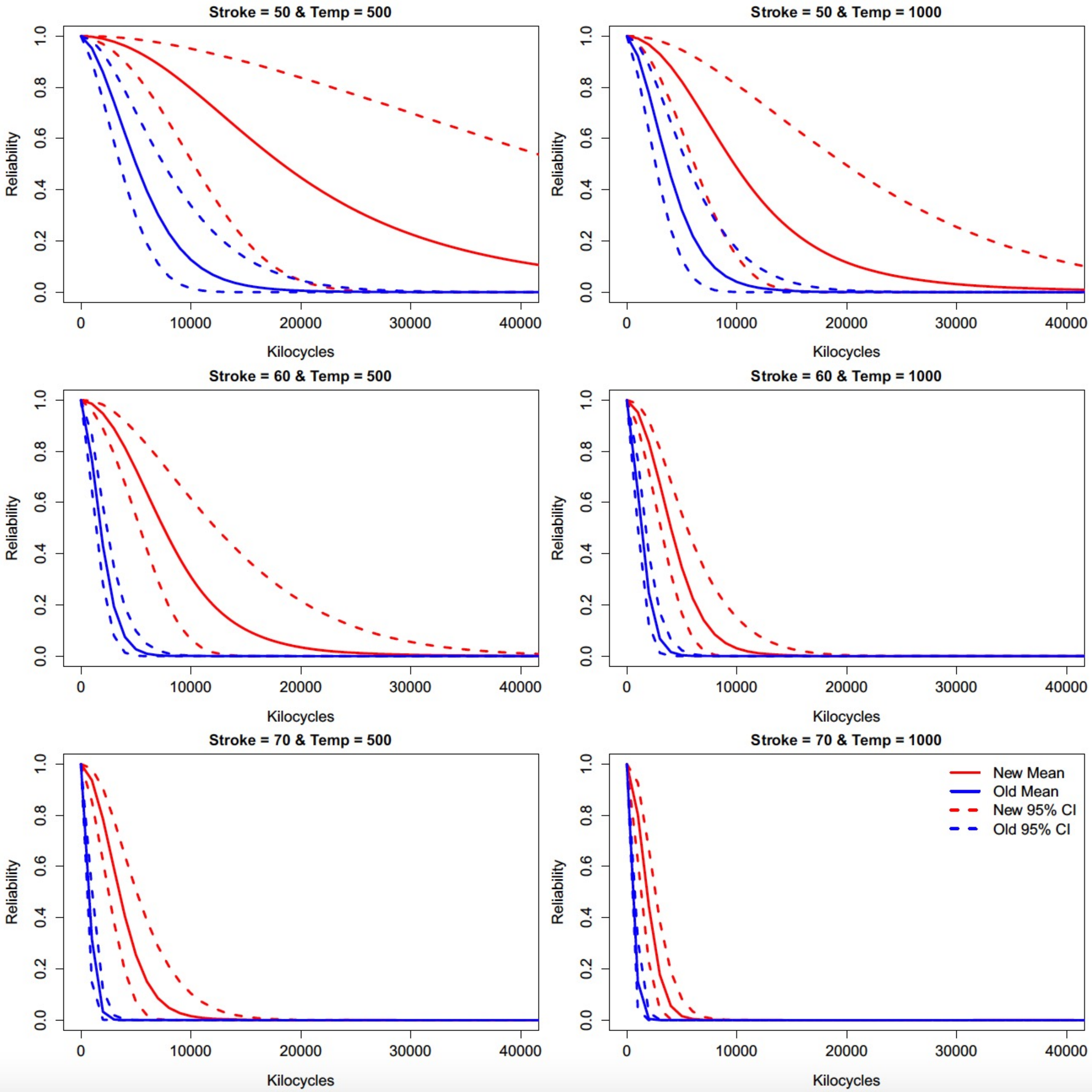
In addition to the processing method, two other design factors were considered and tested at different stress levels.

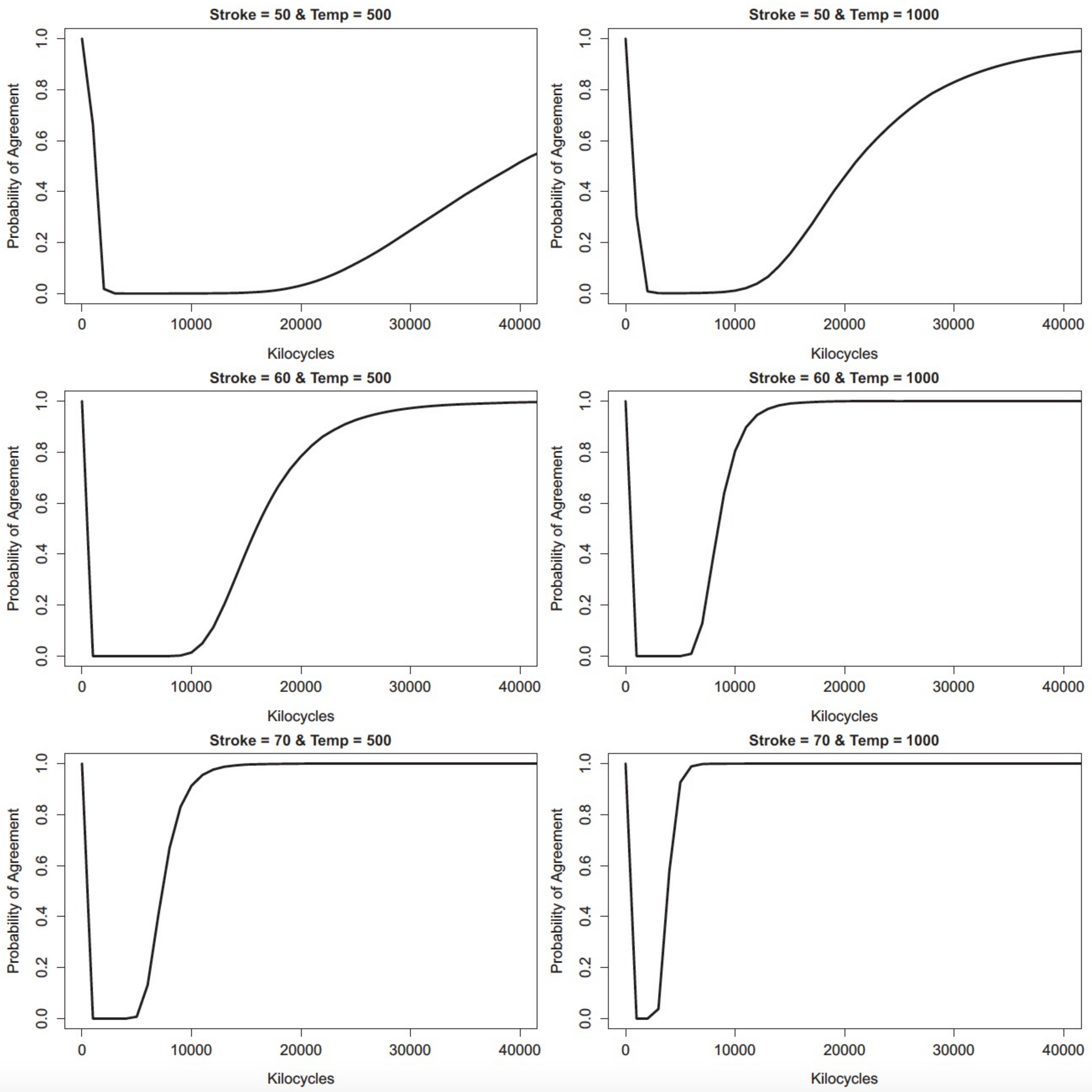
- Process Temperature {500,1000} °F
- Stroke Displacement {50, 60, 70} mils

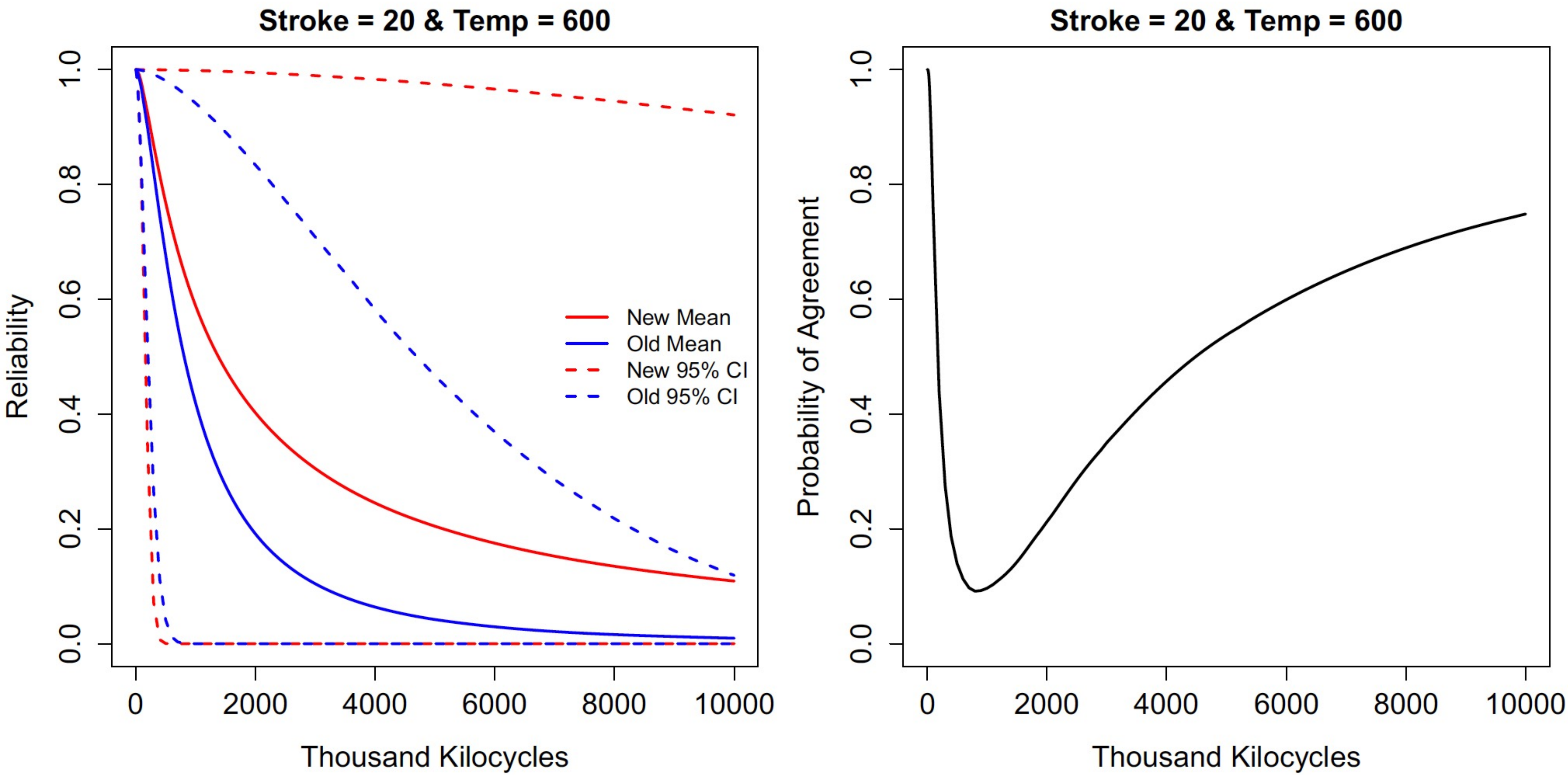
9 replicates were performed at each factorial combination of these factors' levels.

A Weibull distribution was used to model the fatigue life of the spring, with stroke and temperature included as covariates.

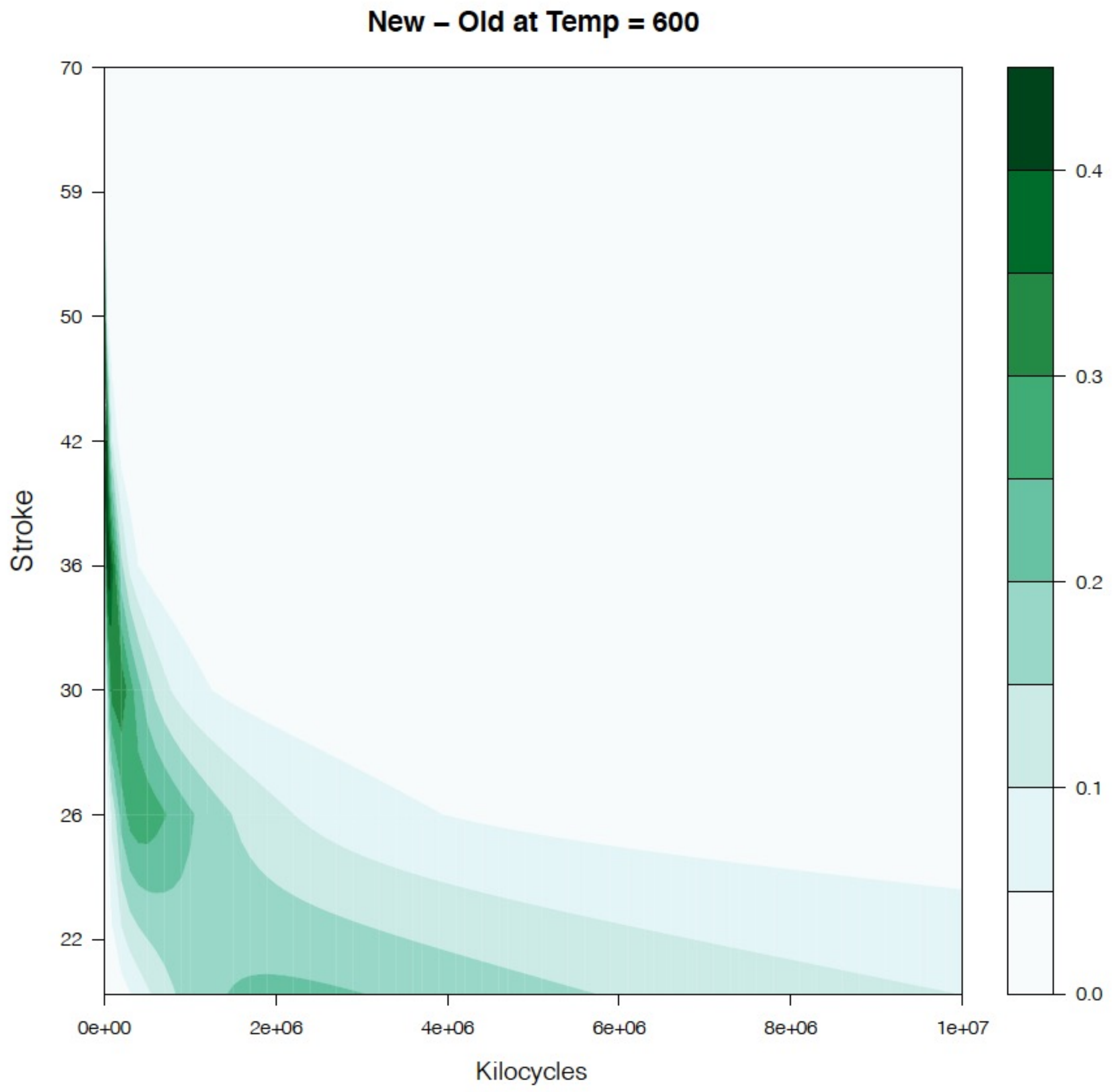


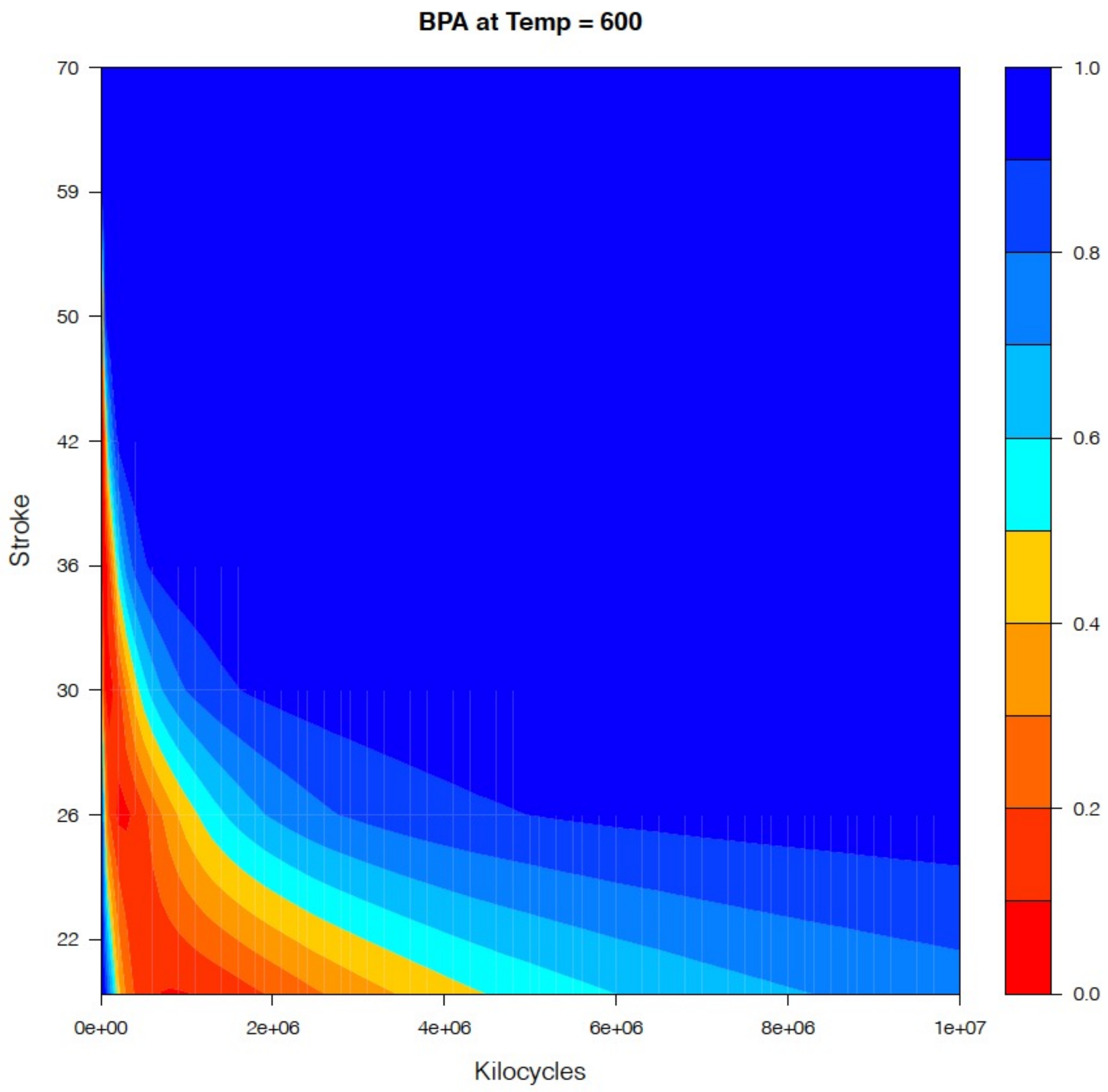

Stroke = 50 & Temp = 500
Stroke = 50 & Temp = 1000
Stroke = 60 & Temp = 500
Stroke = 60 & Temp = 1000
Stroke = 70 & Temp = 500
Stroke = 70 & Temp = 1000
Reliability
Kilocycles
0.0
0.2
0.4
0.6
0.8
1.0
0
10000
20000
30000
40000
New Mean
Old Mean
New 95% CI
Old 95% CI


Stroke = 50 & Temp = 500
Stroke = 50 & Temp = 1000
Stroke = 60 & Temp = 500
Stroke = 60 & Temp = 1000
Stroke = 70 & Temp = 500
Stroke = 70 & Temp = 1000
Probability of Agreement
Kilocycles
0.0
0.2
0.4
0.6
0.8
1.0
0
10000
20000
30000
40000



Stroke = 20 & Temp = 600
Reliability
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000
Thousand Kilocycles
New Mean
Old Mean
New 95% CI
Old 95% CI
Stroke = 20 & Temp = 600
Probability of Agreement
Thousand Kilocycles


New – Old at Temp = 600
Stroke
70
59
50
42
36
30
26
22
Kilocycles
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
0.4
0.3
0.2
0.1
0.0


BPA at Temp = 600
Stroke
70
59
50
42
36
30
26
22
Kilocycles
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
1.0
0.8
0.6
0.4
0.2
0.0

Summary

Summary

- The Bayesian probability of agreement provides an intuitive and practically useful means of comparing reliabilities in two populations
  - It directly quantifies the likelihood that the reliabilities (or other functions of the lifetime distributions) are practically equivalent

- 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 BPA is large enough?

Try it Yourself!

https://nathaniel-t-stevens.shinyapps.io/BPA_Lifetime_app/

THANK YOU!

References

1. Nelson, W.B. (1984). *Accelerated testing: statistical models, test plans, and data analysis*. John Wiley & Sons, Inc.
2. Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*, 2nd ed. Chapman and Hall/CRC Press.
3. Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, 73:sup1, 1-19.
4. Stevens N.T., Steiner S.H. and MacKay R.J. (2017). Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. *Statistical Methods in Medical Research*, 26(6): 2487–2504.
5. Stevens N.T., Steiner S.H. and MacKay R.J. (2018). Comparing heteroscedastic measurement systems with the probability of agreement. *Statistical Methods in Medical Research*, 27(11): 3420–3435.
6. Stevens N.T., Rigdon S.E., and Anderson-Cook C.M. (2020). Bayesian probability of agreement for comparing the similarity of response surfaces. *Journal of Quality Technology* 52(1): 67–80.

References

7. Stevens N.T., Rigdon S.E. and Anderson-Cook C.M. (2018). Bayesian probability of predictive agreement for comparing the outcome of two separate regressions. *Quality and Reliability Engineering International*, 34(6): 968–978.
8. Stevens N.T. and Anderson-Cook C.M. (2019). Design and analysis of confirmation experiments. *Journal of Quality Technology* 51(2): 109–124.
9. Meeker, W. Q., Escobar, L. A., & Zayac, S. A. (2003). Use of sensitivity analysis to assess the effect of model uncertainty in analyzing accelerated life test data. *Case studies in reliability and maintenance*, 135-162.