Applications of the Fractional-Random-Weight Bootstrap

Chris Gotwalt, JMP

Yili Hong, Virginia Tech

William Q. Meeker, Iowa State University

Abstract

The bootstrap, based on resampling, has, for several decades, been a widely used method for computing trustworthy confidence intervals for applications where no exact method is available and when sample sizes are not large enough to be able to rely on easy-to-compute large-sample approximate methods, such a Wald (normal-approximation) confidence intervals. There are, however, many applications where the resampling bootstrap method cannot be used. These include situations where the data are heavily censored, logistic regression when the “success” response is a rare event or where there is insufficient mixing of successes and failures across the explanatory variable(s), and designed experiments where the number of parameters is close to the number of observations. The thing that these three situations have in common is that there may be a substantial proportion of the resamples where is not possible to estimate all of the parameters in the model. This paper reviews the fractional-weight r random weight bootstrap method and shows how it can be used to avoid these problems and provide trustworthy confidence intervals.

Keywords: Censored data, Confidence interval, Designed experiment, Prediction interval, Resampling, Variable selection

Contents

1. Introduction

1.1 Bootstrap Background

The bootstrap is a popular statistical tool that is used, primarily, to obtain improved inferences such as approximate confidence intervals and approximate prediction intervals that have coverage probabilities that are close to the nominal confidence level. With modern computing technology (hardware and software) bootstrap methods are easy to implement and use. Bootstrap procedures can be applied even in situations where classical theory offers little or no guidance about how to compute trustworthy confidence interval. Generally, there are only minimal regularity conditions (such as a finite variance and a certain degree of smoothness) that are needed to the methods to work well. Technical details can be found in Hall (1992) and Shao and Tu (1995).

There are many different types of bootstrap procedures which can be broadly partitioned into nonparametric and parametric. Nonparametric bootstrap procedures require no assumptions about the shape underlying data-generating probability distribution. Most bootstrap procedures do assume sampling from a continuous or approximately continuous distribution.

Bootstrapping is most commonly done via a Monte Carlo simulation. The most common approach is to generate a sequence of new data sets using "resampling." In resampling, each new data set is generated by sampling the rows of the original data with replacement. This approach for generating the bootstrap samples is nonparametric because no assumption about the shape of the underlying distribution is required.

Bootstrap samples can also be generated by assuming a particular parametric distribution and simulating from that distribution. In applications where censoring or truncation is involved, censoring and truncation must be done in a manner that mimics what was done in the original data-generating process. For example, if censoring is random, then a model for the censoring variable needs to be used in the parametric simulation. Often details about how data were censored are not known or complicated and in such situations, the nonparametric resampling method is easier to implement.

After each bootstrap data set is generated, the statistical procedure (e.g., model fitting and computation of point estimates and in some cases standard errors) is applied to the bootstrap dataset and results are stored. This bootstrap-sample generation/estimation procedure is repeated a number of times (e.g., 2,500 times) and then the saved results are processed to make inferences (e.g., construct confidence intervals). There are many different ways to use bootstrap samples to compute a confidence interval (e.g., simple percentile, bias-corrected percentile, BCA, percentile-*t* etc.).

1.2 Overview

The remainder of this paper is organized as follows. Section 2 weighted data and weighted estimation, concepts that are needed to use the fractional-random-weight (FRW) bootstrap and to see the connection with the traditional resampling method of generating bootstrap samples. Section 3 gives a simple example that compares the traditional resampling and FRW bootstrap methods of generating bootstrap samples, provides some background and motivation for the FRW bootstrap and reviews relevant literature. Section 4 provides an example, involving censored field-failure data where the resampling method encounters difficulties and should be avoided but where the FRW bootstrap works well. Section 5 provides an example involving current-status censored data where again, the resampling method fails but the FRW bootstrap works well. Section 6 describes an application to predict the lifetime of power transformers. Section 7 describes an example where, even though there was no censoring, the resampling bootstrap failed to work correctly due to sensitivity in estimating the shape parameter of the generalized gamma distribution. Section 8 describes an example where the FRW bootstrap is used to find an appropriate model to describe the results of a designed experiment. Section 9 provides some concluding remarks and areas for further research.

2. Weighted Data and Weighted Estimation

2.1 Weighted data

In many data analysis applications, it is convenient to put "weights" (or frequencies or counts) on observations. For example, binary data such as 0010001000100010001 are usually replaced with counts of the number of zeros and ones. Weights are frequently used in life test data. For example, the data typically consist of the failure times (which all have weight 1 unless there are ties---often caused by failures being recorded at discrete inspection times) plus the number of units that survived a 1,000-hour test). Weights are also used in data compression where data are "binned" and the weights indicate the number of observations in each bin (e.g., as displayed in a histogram). When observations have known non-constant variances, it is appropriate to use weights that are inversely proportional to the variance of each observation. The resampling bootstrap method of generating bootstrap samples can also be viewed as data with random with integer weights. That is each observation has a weight indicating the number of times it was drawn in the resample. We will give an explicit example of this in the next section.

2.2 Estimation with weighted data

Many statistical estimation methods allow the specification of weights or frequencies. For example, suppose that a data set consists of $x_1, x_2, \ldots, x_n$ with weights $\omega_1, \omega_2, \ldots, \omega_n$. Then estimates of the mean and standard deviation can be computed from

$$\widehat{\mu}^* = \frac{1}{\sum_{i=1}^{n} \omega_i} \sum_{i=1}^{n} \omega_i x_i \quad \text{and} \quad \widehat{\sigma}^* = \left[\frac{1}{\sum_{i=1}^{n} \omega_i} \sum_{i=1}^{n} \omega_i (x_i - \widehat{\mu}^*)^2 \right]^{1/2}$$

There are similar equations for more general weighted least squares for linear regression models. More generally, suppose we have a dataset $\text{data}_1, \text{data}_2, \ldots, \text{data}_n$ with corresponding weights $\omega_1, \omega_2, \ldots, \omega_n$ where each data_i may contain information such as a response, explanatory variables and censoring or truncation indicators. Then the weighted likelihood

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \text{DATA}) = \mathcal{C} \prod_{i=1}^{n} [L_i(\boldsymbol{\theta}; \text{data}_i)]^{\omega_i}$$

3. An Example Illustrating and Comparing Integer and Fractional-Weight Bootstrap Samples

3.1 Integer-weight bootstrap samples

The first column of Table 1gives data tree volume for 15 loblolly pine trees in units of cubic meters. The data are a subsample of data analyzed in Chapter 13 of Meeker Hahn and Escobar (2017). The other three columns give the results of resampling with replacement from the sample of size 15, indicating the number of times that each tree was selected for each of the three resamples. As described in Section 1, in an actual application of the bootstrap the resampling would be done B times, usually on the order of thousands. Then a weighted estimation method could be applied to each bootstrap resample to obtain the *B* bootstrap estimates. This resampling is equivalent to choosing the weights from a uniform multinomial distribution (i.e., a multinomial distribution with uniform probability $1/n$ for each of the original observations in the sample). The uniform multinomial distribution has a mean 1 and a variance $(n-1)/n$.

Table 1 Three Bootstrap Resamples Displayed Using Integer Weights

	Uniform Multinomial distribution Integer weights		
Tree volume	$j=1$	$j=2$	$j=3$
0.149	1	1	1
0.086	2	0	0
0.149	3	0	0
0.194	0	0	1
0.044	1	1	0
0.104	1	1	1
0.156	0	2	1
0.122	1	0	1
0.117	0	3	2
0.079	3	0	2
0.179	0	0	1
0.307	0	7	0
0.049	0	0	1
0.165	1	0	2
0.043	2	0	2
	15	15	15

3.2 Fractional-weight bootstrap samples

Table 2 is similar to Table 1 but has three additional columns containing random fractional weights drawn from a uniform Dirichlet distribution, multiplied by n. These weights, like the integer multinomial resampling weights, will sum to n n), have expectation 1, but a variance $(n-1)/(n+1)$.

The FRW bootstrap was first suggested by Rubin (1981). He called it the "Bayesian bootstrap" because, as shown in the paper, estimates computed from the FRW bootstrap samples are draws from a posterior distribution under a particular relatively diffuse prior distribution. Newton and Raftery (1994) apply Rubin's ideas to a sequence of parametric inference examples. Even though these random-weight bootstrap methods were developed within a nonparametric Bayesian framework, they also apply to non-Bayesian and parametric inference problems, as will be illustrated in the examples in this paper.

There are statistically valid alternative methods to generate the random fractional weights. In particular, Jin et al. (2001) show that FRW bootstrap estimators have good properties if positive independent and identically distributed weights are generated from a continuous distribution that has a mean and standard deviation equal to one (e.g., an exponential distribution with mean one).

Table 2 Three Integer-Weight (on the Left) and Fractional-Weight (on the Right) Bootstrap Samples

	Uniform Multinomial distribution Integer weights			Uniform Dirichlet distribution Continuous weights		
Tree volume	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
0.149	1	1	1	0.203	0.485	1.451
0.086	2	0	0	0.065	1.328	2.062
0.149	3	0	0	0.629	1.737	0.676
0.194	0	0	1	0.505	0.953	0.590
0.044	1	1	0	0.735	1.510	0.580
0.104	1	1	1	2.543	0.320	2.512
0.156	0	2	1	2.650	0.714	1.320
0.122	1	0	1	0.690	2.072	0.650
0.117	0	3	2	1.095	0.017	0.901
0.079	3	0	2	2.075	1.344	0.792
0.179	0	0	1	0.020	2.368	0.061
0.307	0	7	0	1.947	0.116	1.917
0.049	0	0	1	1.433	0.633	0.982
0.165	1	0	2	0.131	1.137	0.212
0.043	2	0	2	0.279	0.265	0.294
	15	15	15	15	15	15

3.3 Fractional-random-weight bootstrap background

The FRW or Bayesian bootstrap is also known as the

- Random-weight bootstrap
- Weighted likelihood bootstrap
- Weighted bootstrap
- Perturbation bootstrap

Operationally, the FRW bootstrap samples are used in the same way as the resampling bootstrap samples. Like resampling, the method is nonparametric. There are, however, important advantages of using the FRW bootstrap in certain common applications. The advantages arise because all of the original observations remain in all of the bootstrap samples. In situations where dropping certain observations from a data set will cause estimation problems, the resampling bootstrap sampling approach will often give poor results or fail altogether. Generally, when using the FRW bootstrap, because all of the original observations remain in the sample, estimation difficulties do not arise.

3.4 Literature review

Much has been written about the bootstrap methods since their introduction in the late 1970s. For example, the textbooks by Efron and Tibshirani (1993), Davison and Hinkley (1997) describe bootstrap theory and methods. The books by Hall (1992) and Shao and Tu (1995) focus on the theory behind bootstrap methods. Another notable reference, aimed at teaching bootstrap methods, is Hesterberg (2015)

In spite of its importance, the FRW bootstrap sampling method does not seem to be well known and there is not much literature using or describing the method. Barbe and Bertail (1995) provide a highly technical presentation of the asymptotic theory of various random-weight methods for generating bootstrap estimates. They show how to choose the distribution of the random weights by using Edgeworth expansions. Chatterjee and Bose (2005) present a generalized bootstrap for which the traditional resampling and various weighted likelihood and other weighted estimating equation methods are special cases. Chiang et al. (2005) apply the FRW bootstrap methods to a recurrent events application with informative censoring in a semi-parametric model. Hong et al. (2009) apply FRW bootstrap methods to a prediction interval application involving complicated censoring and truncation. Xu, Hong, and Meeker (2015) use the FRW bootstrap in a prediction application to assess the risk of future failures of a serious failure mode.

4. Bearing Cage Field Failure Data

4.1 Background

There were 1703 aircraft engines that had been put into service over time, as shown in the event plot in Figure 1. There had been 6 failures and there were 1697 right-censored observations. These data were originally given in Abernethy et al. (1983) and were re-analyzed in Chapter 8 of Meeker and Escobar (1998).

4.2 Weibull analysis

Figure 2 is a Weibull probability plot of the field-failure data. Figure 3 summarizes the numerical results of the estimation. For this example, we will focus on the estimation of the Weibull shape parameter β. The maximum likelihood (ML) point estimate is 2.035. The Wald upper endpoint of the 95% confidence interval is 5.67. Because of the small number of failures, the Wald confidence interval is not trustworthy. JMP also provides likelihood-based confidence intervals. The likelihood upper endpoint (details not shown here) is 3.58. Another alternative for computing trustworthy confidence intervals is the bootstrap. Care is needed, however, when using the resampling bootstrap method with heavy censoring. If the expected number failing is too small there could be bootstrap samples with only 0 or 1 failures, causing JMP's ML algorithm to fail. For the Bearing Cage example, the probability of obtaining a bootstrap sample with 0 or 1 failures using the resampling method is (based on a simple binomial distribution computation) 0.017. Using the FRW method, the probability is 0!

Event Plot

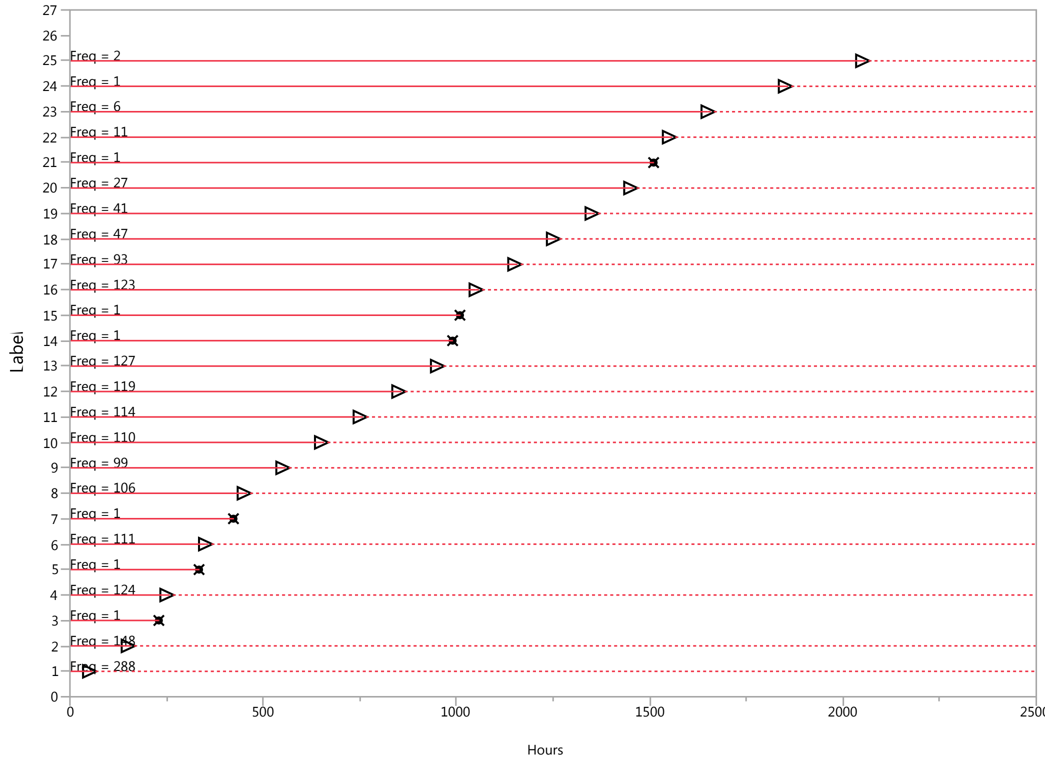


Figure 1 JMP Event Plot for the Bearing Cage Field-Failure Data

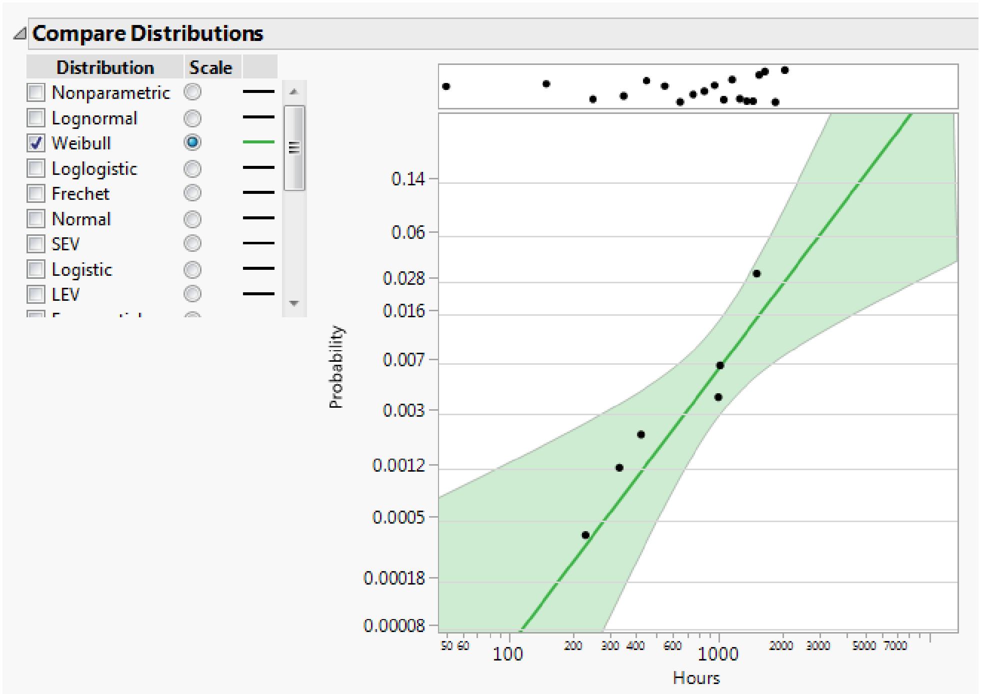


Figure 2 JMP Weibull Probability Plot of the Bearing Cage Field-Failure Data

Parametric Estimate - Weibull

Parameter	Estimate	Std Error	Lower 95%	Upper 95%
location	9.375	0.8351	7.7383	11.012
scale	0.491	0.1607	0.1764	0.806
Weibull α	11792.178	9848.1267	2294.6744	60599.215
Weibull β	2.035	0.6657	1.2403	5.670
Mean	10447.606	8771.4265	2015.4834	54156.970

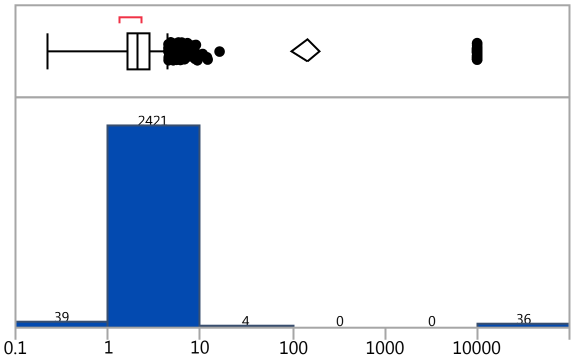
Criterion	
-2*LogLikelihood	152.87379
AICc	156.88085
BIC	167.75409

Figure 3 JMP estimates for the Weibull Analysis of the Ball Bearing Life Test Data

4.3 Bootstrap results

Figure 4 shows results from the resampling (left) and FRW bootstrap distribution for the Weibull shape parameter β. The histogram on the left shows that there were 36 samples that resulted in a wild estimate of β which were probably caused by having resamples with 0 or 1 failures. The upper endpoint of the confidence interval is even more extreme than that provided by the untrustworthy Wald method. The histogram on the right, based on the FRW bootstrap method is better behaved and the upper endpoint only 4.4. This is a more trustworthy value and is consistent with common experience with fatigue failures in the field. Interestingly (but not surprisingly) the FRW method runs somewhat faster than the resampling method for this example. This is because with the FRW method the optimization algorithm is not faced with bootstrap samples that result in poorly behaved likelihoods that require a lot of time trying to find the maximum.

Weibull β

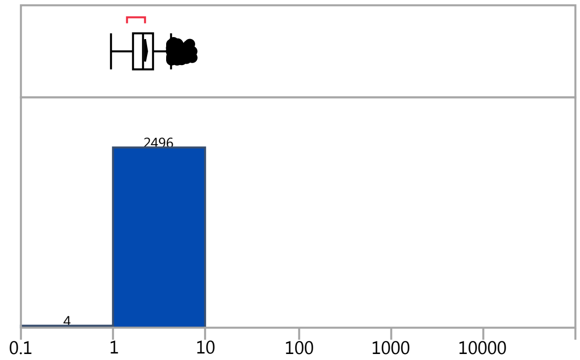


Bootstrap Confidence Limits

Coverage	BC Lower	BC Upper
0.95	1.03509	6.54006
0.90	1.15301	4.86636
0.80	1.30372	3.74995
0.50	1.57374	2.6705

Original Estimate 2.035319

Weibull β



Bootstrap Confidence Limits

Coverage	BC Lower	BC Upper
0.95	1.1882	4.40162
0.90	1.26962	3.89811
0.80	1.38419	3.34357
0.50	1.62542	2.64119

Original Estimate 2.035319

Figure 4 Resampling (left) and Fractional-Random-Weight (right) Bootstrap Results for the Weibull Shape Parameter for the Bearing Cage Field-Failure Data

5. Rocket Motor Field-Failure Weibull Analysis

5.1 Background

Olwell and. Sorell (2001) present a Bayesian analysis of rocket motor field-failure data. The data were reanalyzed in Chapters 14 and 18 of Meeker, Hahn, and Escobar (2017). There were approximately 20,000 missiles in inventory that had been manufactured over a number of years and put into the stockpile. There had been 1,940 rockets put into flight over a period of time up to 18 years subsequent to their manufacture. At their time of flight, 1,937 of these motors performed satisfactorily; but there were three catastrophic launch failures. The failure probability at 20 years was of interest.

Table 3 Rocket Motor Life Data (in Years Since Manufacture).

Years	Number of motors	Years	Number of motors	Years	Number of motors
> 1	105	> 8	211	> 14	14
> 2	164	> 9	124	> 15	5
> 3	153	> 10	90	> 16	3
> 4	236	> 11	72	< 8.5	1
> 5	250	> 12	53	< 14.2	1
> 6	197	> 13	30	< 16.5	1
> 7	230				

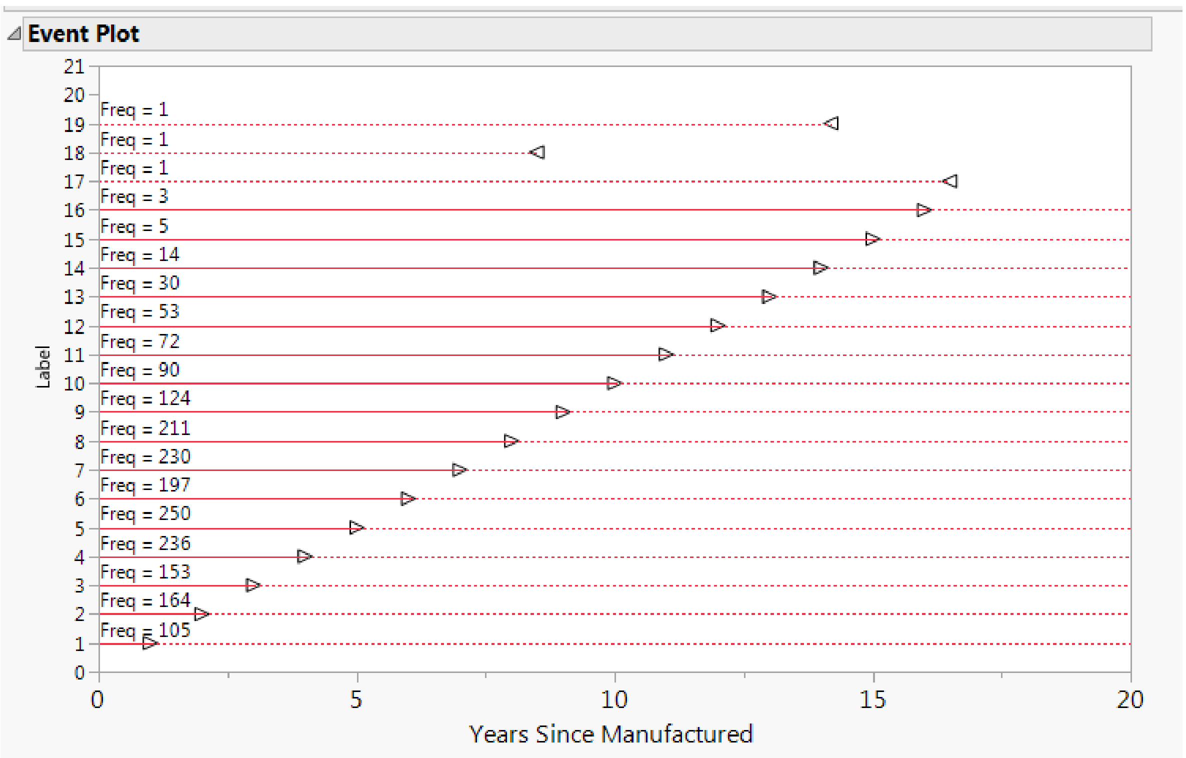


Figure 5 JMP Event Plot for the Rocket Motor Field-Failure Data

The data are shown in Table 3 where > indicates that the failure time was greater that the indicated year (right-censored observations) and < indicates that the failure time was less than the indicated year (left censored observations). Figure 5 is an event plot showing the structure of the data. Although these data could be described with a binary regression model (failure probability as a function of years since manufactured), there are advantages of treating such data as failure-time data. In particular, the fraction failing as a function of time is constrained to be monotone increasing and we can use probability plotting to assess whether a chosen distribution is appropriate.

The usual resampling (integer weight) bootstrap will not work well with this data set, as demonstrated below.

5.2 Weibull analysis

Figure 6 is a Weibull probability plot of the rocket motor field failure data. Table 4 gives the numerical ML results and Wald confidence intervals for the distribution parameters. As with the bearing cage example, we will again focus on the Weibull shape parameter to compare the different confidence intervals. The upper endpoint of the Wald confidence interval is 34.6. JMP can also provide the likelihood interval which has an upper endpoint of 15.54, considerably smaller than then that for the Wald interval. Generally, the likelihood interval would be considered to be more trustworthy than the Wald interval. The bootstrap interval might also be expected to provide a trustworthy interval. In this case, however, because of the heavy censoring, the resampling method will not work properly.

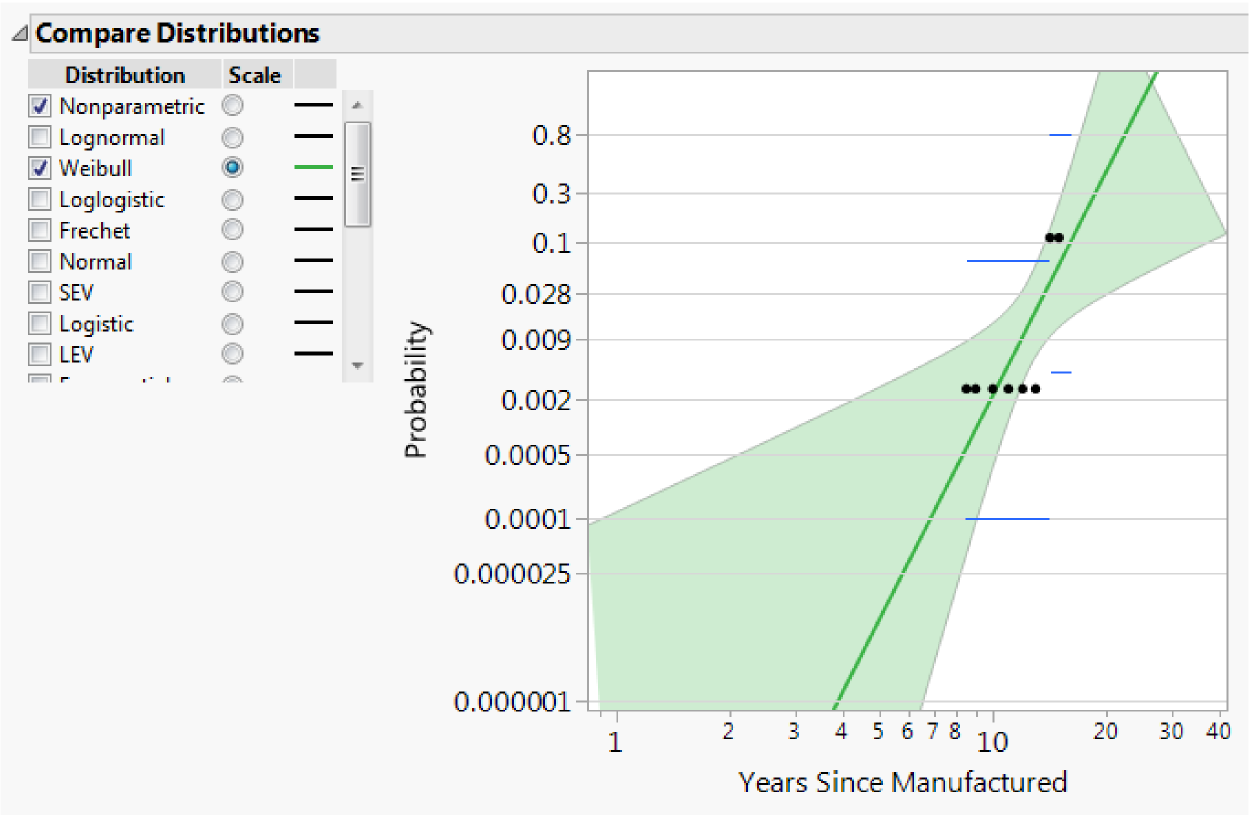


Figure 6 JMP Weibull Probability Plot of the Rocket Motor Field-Failure Data

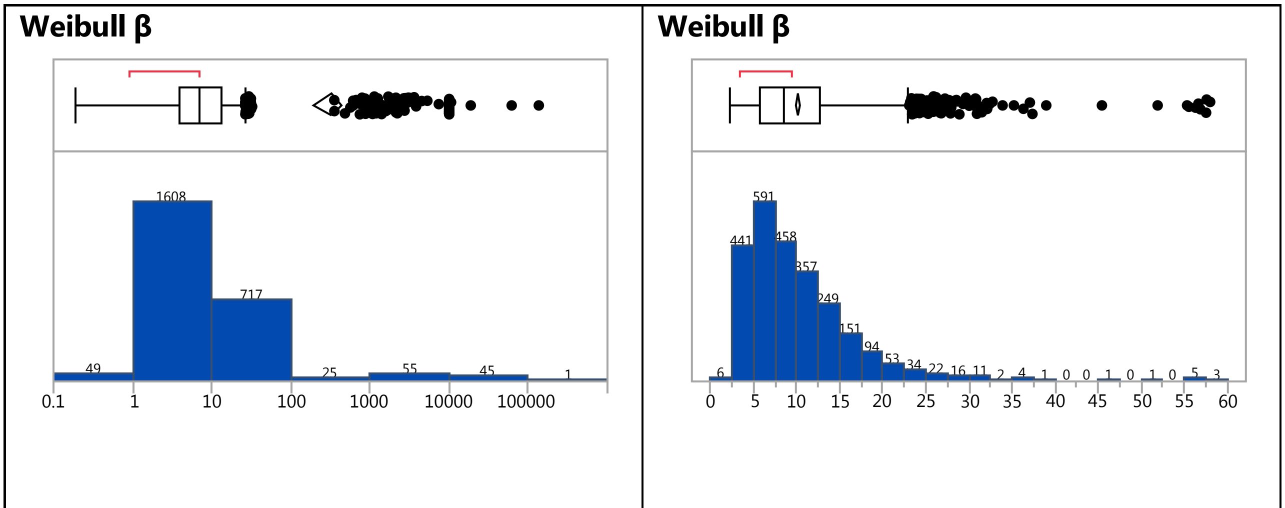
Table 4 ML Weibull Estimation Results and Wald Confidence Intervals for the Rocket Motor Field Failure Data

Parametric Estimate - Weibull

Parameter	Estimate	Std Error	Lower 95%	Upper 95%
location	3.055358	0.2162706	2.824124	4.210590
scale	0.123058	0.0480368	0.064348	0.337550
Weibull α	21.228791	4.5911625	16.846186	67.396303
Weibull β	8.126236	3.1721454	2.962524	15.540500
Mean (Wald CI)	20.007199	3.9737989	13.555700	29.529128

5.3 Bootstrap results

Figure 7 shows the results for bootstrapping the ML estimate of the Weibull shape parameter for the rocket motor. Using the resampling method there was a large number of samples giving a very large value of the bootstrap estimate of β. These probably arose from bootstrap samples that had no or only one of the left-censored observations in the bootstrap sample. Using the FRW bootstrap, on the other hand resulted in a much more reasonable distribution and corresponding 95% confidence interval for β.



Bootstrap Confidence Limits (left)

Coverage	BC Lower	BC Upper
0.95	1	10000
0.90	1.20999	2455.21
0.80	2.29759	30.227
0.50	5.3482	19.4609

Original Estimate 8.126236

Bootstrap Confidence Limits (right)

Coverage	BC Lower	BC Upper
0.95	2.83205	22.912
0.90	3.09711	19.3049
0.80	3.66637	15.9875
0.50	5.20738	11.4789

Original Estimate 8.126236

Figure 7 Resampling (left) and Fractional-Random-Weight (right) Bootstrap Results for the Weibull Shape Parameter for the Rocket Motor Field-Failure Data

6. Bootstrapping a Model for the Power Transformer Field Data

6.1 Background

Extending previous work of Escobar and Meeker (1999) and Lawless and Fredette (2005), Hong, Meeker, and McCalley (2009) describe the use of the FRWed bootstrap to generate prediction intervals for the number of power transformers that will need to be replaced in future years. The dataset contained information on 710 power transformers with 62 units that had failed. Units still in service at the "data freeze" date in March 2008 are right censored. Some units that were still in service were more than 60 years old. One difficulty with the data is that records of transformers removed from service before 1980 were not available. Thus units that had been installed before 1980 and which were still in service are observations from a truncated distribution. Figure 8 is an event plot of a representative subset of the data. The red lines with circles on the right indicate transformer age at the time of failure. The black lines indicated units that were still in service at the time of the analysis.

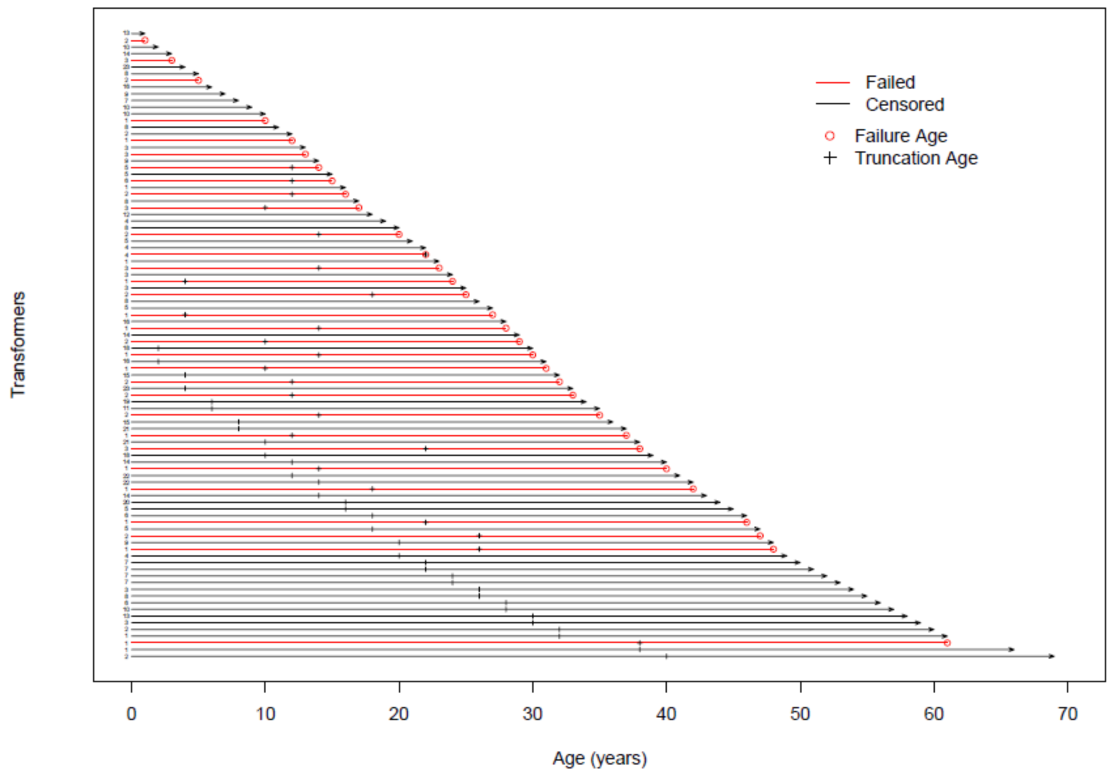


Figure 8 Event Plot for the Power Transformer Field-Failure Data

There were several covariates, including manufacturer and cooling method that had an effect on the life distribution. It was also learned that, even after adjustment for the other covariates, that there was an important difference between the failure-time distributions of transformers manufactured before and after the mid-1980s. Transformers manufactured before the mid-1980s tend to have longer lifetimes, due to over-engineering that was practiced then.

6.2 Modeling and maximum likelihood estimation

Weibull and lognormal distributions were fit to the data using maximum likelihood. Stratification was based on whether manufactured before or after 1987. The likelihood functions is

$$L(\boldsymbol{\theta}|DATA) = \prod_{i,j} f(t_{ij};\boldsymbol{\theta})^{\delta_{ij}\nu_{ij}} \times \left[\frac{f(t_{ij};\boldsymbol{\theta})}{1-F(\tau_{ij}^{L};\boldsymbol{\theta})}\right]^{\delta_{ij}(1-\nu_{ij})} \times \left[1-F(c_{ij};\boldsymbol{\theta})\right]^{(1-\delta_{ij})\nu_{ij}} \times \left[\frac{1-F(c_{ij};\boldsymbol{\theta})}{1-F(\tau_{ij}^{L};\boldsymbol{\theta})}\right]^{(1-\delta_{ij})(1-\nu_{ij})}$$

where t_{ij} is the failure t_{ij} is the censoring time τ_{ij}^{L} is the lower truncation time and δ_{ij} and ν_{ij} are censoring and truncation indicators for transformer *i* in stratum *j* and the $\boldsymbol{\theta}$ is a vector of parameters where the distribution scale parameter is molded as a function. An important question was how to generate bootstrap samples to do the calibration of the prediction intervals. The commonly-used parametric bootstrap would be complicated to implement because it would require a model for the censoring and truncation processes. The resampling method would also have difficulties because of the categorical variables and the small number of failures in some of the categories. The FRW bootstrap offered an attractive, easy-to-implement alternative that worked without any problems.

6.3 Prediction results

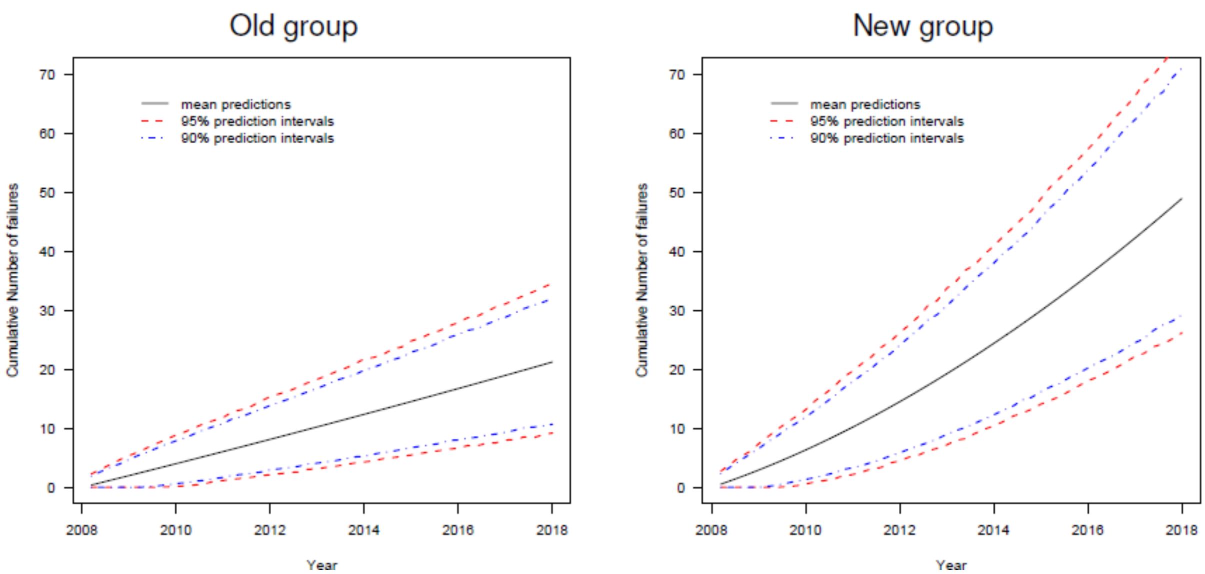


Figure 9 Power Transformer Fleet Predictions Based on the Fractional-Random-Weight Bootstrap

Figure 9 shows predictions and prediction intervals for the cumulative number of transformer failures for the ten years starting in 2008. The plot on the left is for transformers that were installed in 1987 or before. The plot on the right is for transformers that were installed after 1987. The predicted number failing for the latter group is larger both because the lifetimes of the newer models tend to be shorter and also because the risk set was larger.

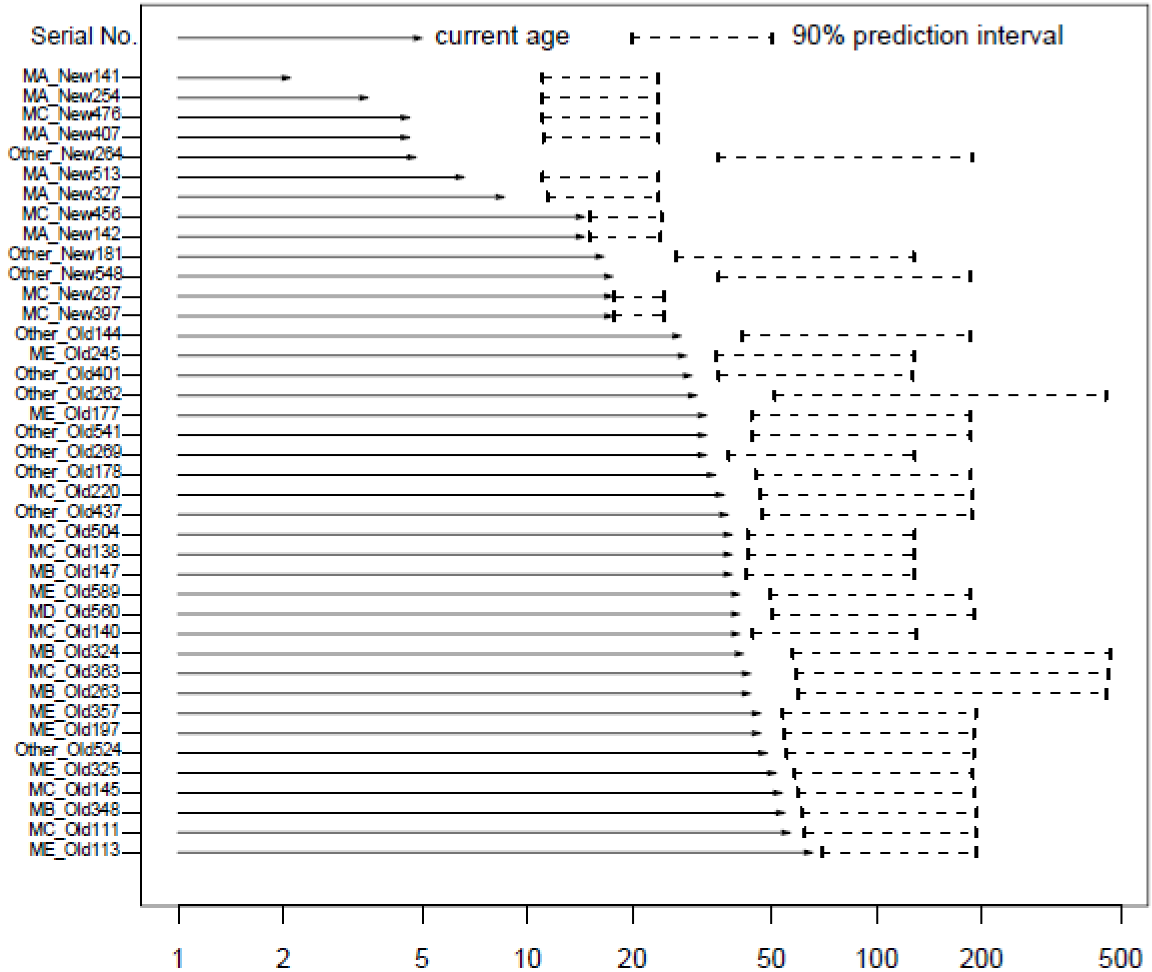


Figure 10 Power Transformer Individual Predictions based on the Fractional-Random-Weight Bootstrap

Figure 10 shows, for a subset of the transformers that were still in operation at the time the predictions were made, the age of the transformer and a prediction interval quantifying the information available about the distribution of remaining life for individual transformers. Although some of the upper endpoints of the prediction intervals may be hard to believe, the lower endpoints allowed a ranking of which transformers where at highest risk for failure in the short term.

7. Bootstrapping the Generalized Gamma Distribution Model for the Ball Bearing Failure Time Data

7.1 Background

Meeker and Escobar (1998) and Lawless (2003) fit the generalized gamma distribution to ball bearing life test data that were originally reported in Lieblein and Zelen (1956). Figure 11 is an event plot of the data. There was no censoring.

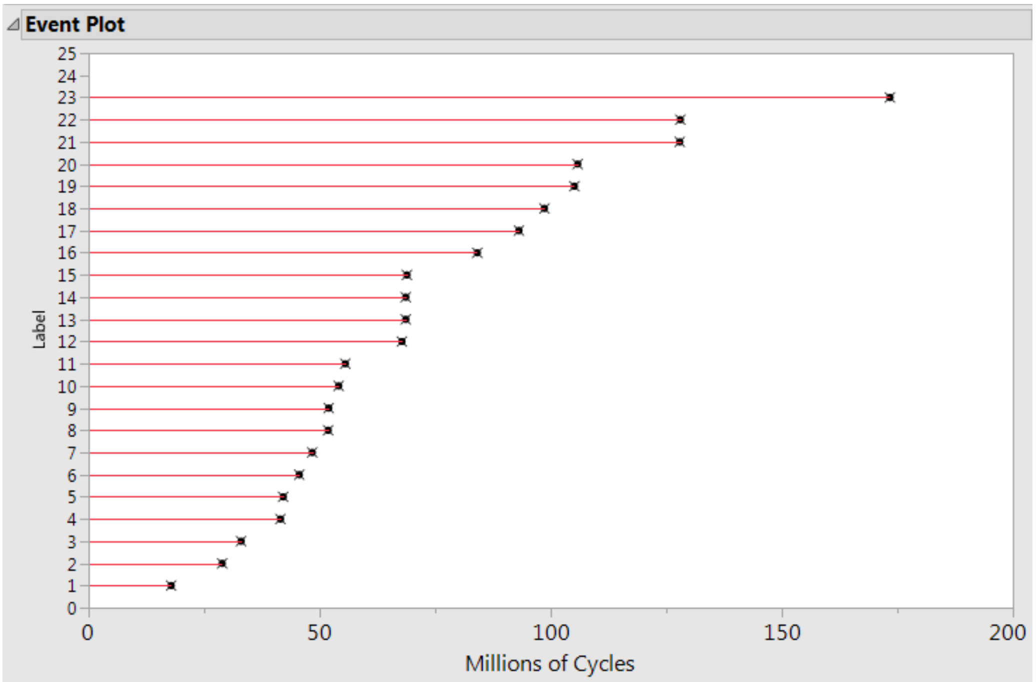


Figure 11 JMP Event Plot for the Ball Bearing Life Test Data

The generalized gamma distribution is interesting in that depending on the value of the shape parameter λ, the Weibull ($\lambda = 1$), lognormal ($\lambda = 0$), and Frechet ($\lambda = -1$) distributions are all special cases.

7.2 Maximum likelihood estimation

Figure 12 shows the generalized gamma distribution ML estimates plotted on Weibull probability paper. We can see that it generally fits the data well. Table 5 gives the ML estimation results and Wald confidence intervals for the parameters. The 95% likelihood-based confidence interval for λ is [-0.76, 1.53}, somewhat wider than the Wald confidence interval. Generally, the likelihood-based interval is more trustworthy. In either case, the confidence interval for λ provides an indication that both the lognormal and Weibull distributions are consistent with the data. This is because both 0 and 1 lie inside the confidence interval.

The bootstrap provides another alternative to computing confidence intervals for this distribution. In the next section, we will compare the resampling and the FRW bootstrap methods.

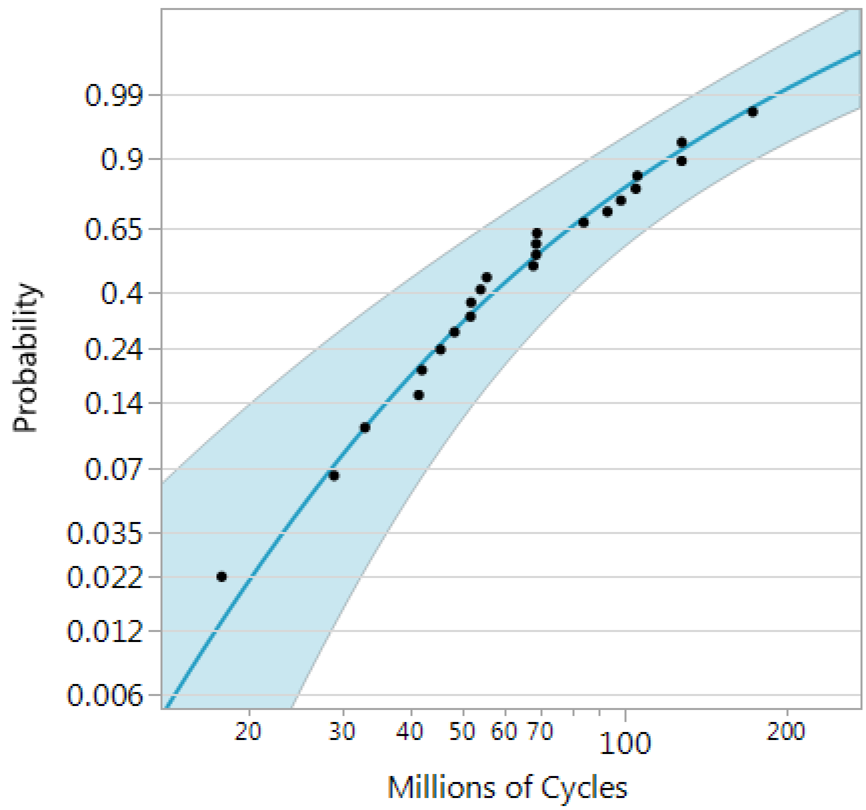


Figure 12 Ball Bearing Failure Times Generalized Gamma Analysis on Weibull Probability Plot

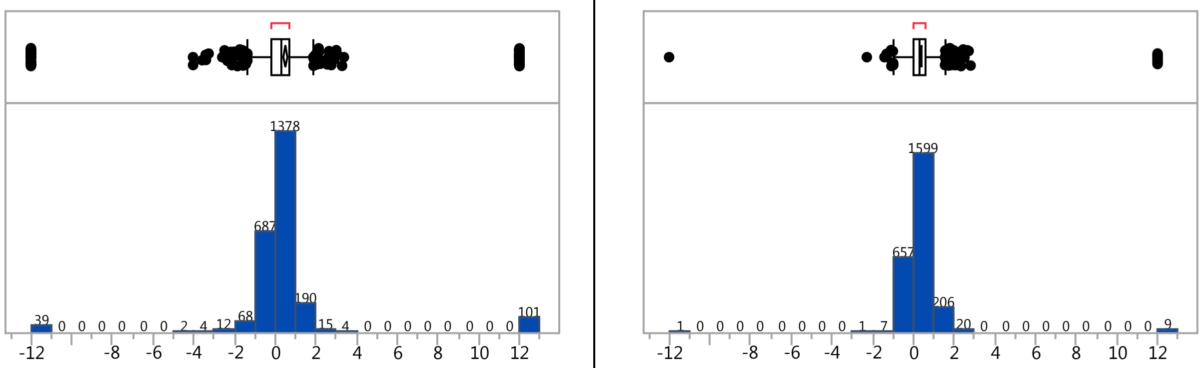
Table 5 ML Generalized Gamma Estimation Results and Wald Confidence Intervals for the Rocket Motor Field Failure Data

Parametric Estimate - Generalized Gamma

Parameter	Estimate	Std Error	Lower 95%	Upper 95%
mu	4.2300656	0.17705333	3.883047	4.5770838
sigma	0.5099753	0.07934656	0.354459	0.6654917
lambda	0.3076610	0.54865342	-0.767680	1.3830019

7.3 Bootstrap results

Figure 13 gives bootstrap results for the ball bearing generalized gamma distribution shape parameter λ. The histogram on the left shows results for the resampling bootstrap method; the histogram on the right shows results using the FRW bootstrap method. With the resampling method, there was ML estimate convergence problems with a substantial number of the bootstrap samples. In particular, the estimate of the shape parameter λ was -12 39 times and $+12$ 101 times (the programmed limits in the JMP software). The 101 values at $+12$ were more than enough to cause the upper endpoint of the 95% bootstrap confidence interval to be $+12$. The corresponding numbers when using the FRW bootstrap method were 1 and 9 and these numbers are so small that they have little or no effect on the confidence interval for λ. The FRWt method provides a better method for computing confidence intervals for the generalized gamma distribution when the sample size is not large.



Bootstrap Confidence Limits (left)

Coverage	BC Lower	BC Upper
0.95	-1.318	12
0.90	-0.8765	12
0.80	-0.5386	1.24298
0.50	-0.0716	0.72872

Original Estimate 0.307661

Bootstrap Confidence Limits (right)

Coverage	BC Lower	BC Upper
0.95	-0.5952	1.70428
0.90	-0.4331	1.34602
0.80	-0.2633	1.0393
0.50	0.01059	0.64171

Original Estimate 0.307661

Figure 13 Ball Bearing L Resampling (left) and Fractional-Random-Weight (right) Bootstrap Results for the Generalized Gamma Distribution Shape Parameter

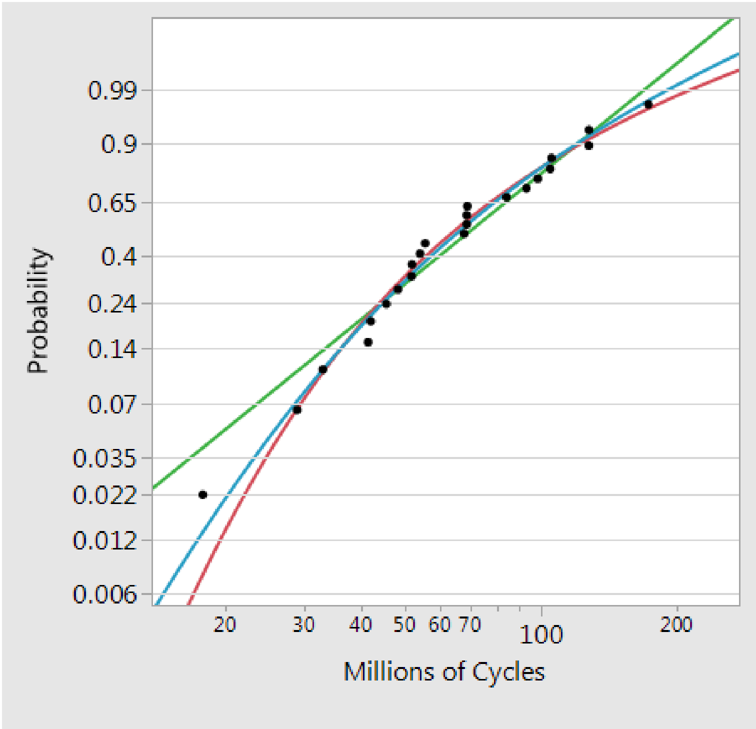


Figure 14 Ball Bearing Failure Times (from top to bottom on the left) Weibull, Generalized Gamma, and Lognormal distributions on Weibull Probability Paper

8. Using Bootstrap to Help Choose a Model for a Designed Experiment

8.1 Background

Designed Experiments (DOEs) are a common approach to problem-solving in science and industry. Experiments are, however, often costly in terms of both time and resources. An often-stated goal is to obtain as much information as possible about the relationship between the experimental factors (x) and response variable(s) (y). Usually, DOEs use a specially constructed combination of x values that optimize information gained in a small number of runs. After the data become available, then there is a need to decide on the appropriate statistical model to describe the relationship between x and y .

8.2 Using the bootstrap in model selection

The bootstrap is a useful tool Identifying the subset of the x variables (as well as possible interaction and quadratic effects) that best explain variation in y . The resampling bootstrap, however, can encounter problems because the removal of observations can drastically change the properties of the

There are two well-known alternatives to resampling: Using a parametric bootstrap (simulating data from a given model) and resampling residuals from a fitted model. The problem with these two methods is that they require specifying a model, which is what we are trying to determine! As mentioned in Section 3.3, the FRW bootstrap is nonparametric and thus suitable for model-building applications that can be used with DOE data.

8.3 The nitrogen oxides example

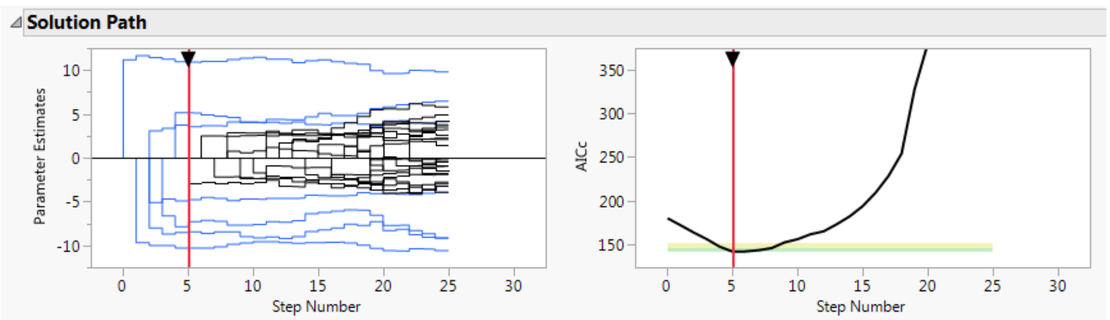
Nitrogen Oxides (NOx) are toxic greenhouse gases that are common by-products of burning organic compounds. An experiment was done on an industrial burner to study the amount of NOx it created. A 32 run I-Optimal RSM design was created with 7 continuous factors:

- Hydrogen Fraction in primary fuel
- Air/Fuel Ratio
- Lance Position X
- Lance Position Y
- Secondary Fuel Fraction
- Dispersant
- Ethanol Percentage in primary fuel

We want to assess the importance of the input variables (including second-order interactions and quadratic terms).

8.4 Using forward selection

First, we apply a forward stepwise procedure that selects a model using the AICc model selection criterion. We did this with the JMP Pro Generalized Regression platform using the Forward Stepwise Selection option. The results are shown in Figure 15.



Parameter Estimates for Original Predictors

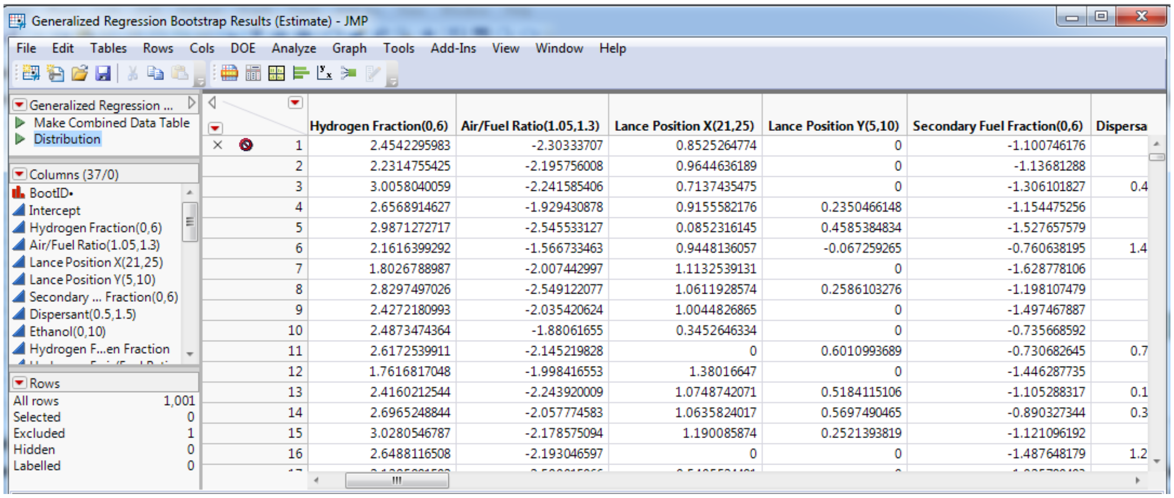
Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%	Singularity Details
Intercept	30.304696	0.431203	4939.1955	<.0001*	29.459553	31.149838	
Hydrogen Fraction(0,6)	2.4542296	0.3412758	51.715329	<.0001*	1.7853412	3.123118	
Air/Fuel Ratio(1.05,1.3)	-2.303337	0.3439153	44.855123	<.0001*	-2.977399	-1.629276	
Lance Position X(21,25)	0.8525265	0.3052632	7.7995036	0.0052*	0.2542217	1.4508313	
Lance Position Y(5,10)	0	0	0	1.0000	0	0	
Secondary Fuel Fraction(0,6)	-1.100746	0.2607764	17.81713	<.0001*	-1.611859	-0.589634	
Dispersant(0.5,1.5)	0	0	0	1.0000	0	0	
Ethanol(0,10)	0	0	0	1.0000	0	0	
Hydrogen Fraction*Hydrogen Fraction	0	0	0	1.0000	0	0	

Figure 15 Results from Using the JMP Pro Generalized Regression Platform to Choose a Model using Forward Stepwise Selection.

To better understand the stability of this model choice and to explore the possibility that other variables might make an important contribution, it is possible to apply the FRW bootstrap method to the model-building procedure. Then the results of such a bootstrap can be used to obtain selection probabilities for the different model terms.

8.5 Bootstrapping the forward selection procedure

Using JMP with the FRW option to bootstrap the forward selection procedure creates a table of the bootstrapped estimates with a Distribution script attached. One thousand FRW bootstrap data sets were generated. For each bootstrap FRW data set, the forward selection procedure is applied and the corresponding row in the table gives the values of the regression coefficients. The zeros in the table indicate that the variable was not included in the model for that bootstrap sample.



	Hydrogen Fraction(0,6)	Air/Fuel Ratio(1.05,1.3)	Lance Position X(21,25)	Lance Position Y(5,10)	Secondary Fuel Fraction(0,6)	Dispersa
1	2.4542295983	-2.30333707	0.8525264774	0	-1.100746176	
2	2.2314755425	-2.195756008	0.9644636189	0	-1.13681288	
3	3.0058040059	-2.241585406	0.7137435475	0	-1.306101827	0.4
4	2.6568914627	-1.929430878	0.9155582176	0.2350466148	-1.154475256	
5	2.9871272717	-2.545533127	0.0852316145	0.4585384834	-1.527657579	
6	2.1616399292	-1.566733463	0.9448136057	-0.067259265	-0.760638195	1.4
7	1.8026788987	-2.007442997	1.1132539131	0	-1.628778106	
8	2.8297497026	-2.549122077	1.0611928574	0.2586103276	-1.198107479	
9	2.4272180993	-2.035420624	1.0044826865	0	-1.497467887	
10	2.4873474364	-1.88061655	0.3452646334	0	-0.735668592	
11	2.6172539911	-2.145219828	0	0.6010993689	-0.730682645	0.7
12	1.7616817048	-1.998416553	1.38016647	0	-1.446287735	
13	2.4160212544	-2.243920009	1.0748742071	0.5184115106	-1.105288317	0.1
14	2.6965248844	-2.057774583	1.0635824017	0.5697490465	-0.890327344	0.3
15	3.0280546787	-2.178575094	1.190085874	0.2521393819	-1.121096192	
16	2.6488116508	-2.193046597	0	0	-1.487648179	1.2

Figure 16 Table of the Bootstrap Estimates with a Distribution Script Attached

Figure 16 shows partial results (i.e., for some of the regression coefficients) from the Distribution script which summarizes all of the bootstrap results. The spikes at 0 in some of the histograms indicate the number of times that the corresponding variable did not enter the model (frequently for Lance Position Y(5,10) and never for Hydrogen Fraction(0,6)).

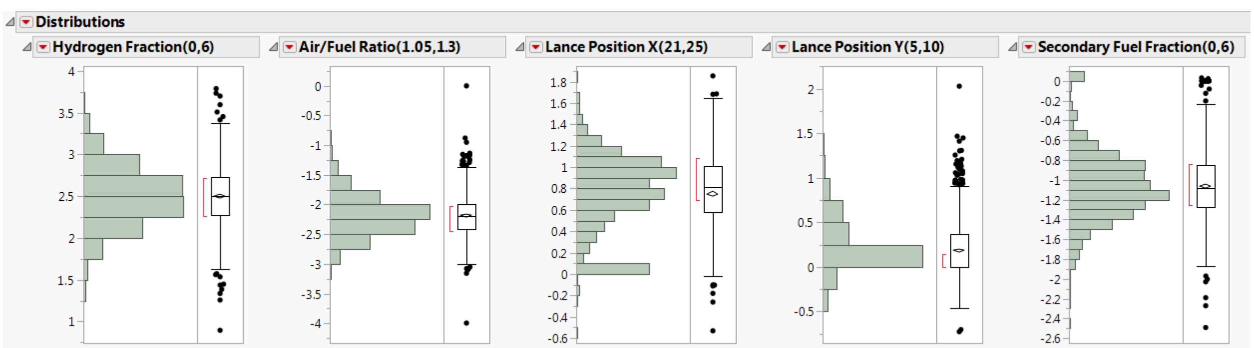


Figure 17 Histograms Created by the Distribution Script Summarize the FRW Bootstrap Modeling Results

The results in Figure 16 can be used to construct a table like that in Figure 18. This table gives the proportion of times across the 1000 bootstrap samples that each variable was chosen to be in the model. Because of the "Bayesian bootstrap" result given in Rubin (1981), these proportions can be interpreted as the posterior probability that the corresponding model term should be in the model. One could then use a cutoff point (such as 0.50) to decide whether model terms should be included or not.

	Term	Proportion Selected
1	Hydrogen Fraction(0,6)	1
2	Air/Fuel Ratio(1.05,1.3)	0.999
3	Secondary Fuel Fraction(0,6)	0.981
4	Air/Fuel Ratio*Air/Fuel Ratio	0.946
5	Lance Position X(21,25)	0.903
6	Lance Position X*Secondary Fuel Fraction	0.848
7	Hydrogen Fraction*Secondary Fuel Fraction	0.733
8	Lance Position Y(5,10)	0.593
9	Dispersant(0.5,1.5)	0.551
10	Secondary Fuel Fraction*Secondary Fuel Fraction	0.479
11	Lance Position Y*Lance Position Y	0.294
12	Hydrogen Fraction*Lance Position Y	0.224
13	Hydrogen Fraction*Hydrogen Fraction	0.203
14	Lance Position Y*Dispersant	0.192
15	Hydrogen Fraction*Lance Position X	0.18
16	Air/Fuel Ratio*Lance Position Y	0.177
17	Air/Fuel Ratio*Dispersant	0.163
18	Ethanol(0,10)	0.113
19	Hydrogen Fraction*Air/Fuel Ratio	0.104
20	Lance Position X*Lance Position X	0.083

Figure 18 The Proportion Of Times Across the 1000 Bootstrap Samples that Each Variable was Chosen to be in the Model

9. Concluding Remarks and Areas for Future Research

With vastly improved computing capabilities and the bootstrap theory that has been developed over the past 40 years, bootstrapping provides an important useful tool for obtaining

- Trustworthy confidence intervals
- Trustworthy prediction intervals
- Better regression models

The FRW bootstrap tremendously expands the potential areas of application of the bootstrap to applications involving heavy censoring and/or truncation, categorical explanatory variable's, and designed experiments where dropping certain combinations of the original observations can cause estimability problems. Those problems do not arise when the FRW bootstrap is used.

There are a number of areas that could be investigated to provide further insight into when and how the FRW bootstrap methods should be used with finite samples.

- As described in Section 3, there are different, asymptotically equivalent ways to choose random weighs for bootstrapping (including resampling). This leaves open the question about differences in the properties of bootstrap procedures in finite samples. For example, if weights are chosen to have a mean and variance of one, what would be the effect on the performance of varying the third or higher moments?

- We have demonstrated a clear advantage for the FRW bootstrap in situations where estimability problems occur when certain combinations of observations are dropped. In situations where there will be no estimability problems is it possible that the FWB approach. It would be useful to compare different nonparametric and parametric methods for generating bootstrap estimates when using a parametric model to describe one's data. In particular, it would be interesting to compare
    - Resampling
    - A fully parametric bootstrap simulation (e.g., where the censoring distribution is modeled).
    - FRW bootstrap

to see if there are important differences in bootstrap performance.

- Generalized Pivotal Quantity (GPQ) inference (also known as Generalized Fiducial inference; see Hannig, Iyer, and Patterson (2006)) has proven to be a powerful tool for defining confidence interval procedures for non-standard models. Implementing GPQ methods generally requires computing a large set of simulated parameter estimates, in a manner that is similar to the parametric bootstrap. In situations involving heavy censoring, even the parametric bootstrap sampling will have estimability problems. Use of FWB instead should allow GPQ methods to be used in such applications.

References

Abernethy, R. B., J. E. Breneman, C. H. Medlin, and G. L. Reinman (1983). *Weibull Analysis Handbook*. Air Force Wright Aeronautical Laboratories Technical Report AFWAL-TR-83-2079. Available from the National Technical Information Service, Washington, DC.

Barbe, P. and P. Bertail (1995). *The Weighted Bootstrap*. Springer.

Chatterjee, S. and A. Bose (2005). Generalized bootstrap for estimating equations. *Annals of Statistics* **33**, 414–436.

Chiang, C.-T., L. F. James, and M.-C. Wang (2005). Random weighted bootstrap method for recurrent events with informative censoring. *Lifetime Data Analysis* **11**, 489–509.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Escobar, L.A., and W.Q. Meeker (1999), Statistical Prediction Based on Censored Life Data. *Technometrics* **41**, 113–124.

Efron, B. and R. Tibshirani (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science 1*, 54–77.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer.

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician 69*, 371–386.

Hannig, J., H. Iyer, and P. Patterson (2006). Fiducial generalized confidence intervals. *Journal of the American Statistical Association 101*, 254–269.

Hong, Y., W. Q. Meeker, and J. D. McCalley (2009). Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *Annals of Applied Statistics* **3,** 857–879.

Jin, Z., Z. Ying, and L. Wei (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88,** 381–390.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data* (Second Edition). John Wiley & Sons, Inc.

Lawless, J. F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* **92** 529–542.

Lieblein, J. and M. Zelen (1956). Statistical investigation of the fatigue life of deep-groove ball bearings. *Journal of Research, National Bureau of Standards* **57**, 273–316.

Meeker, W. Q. and L. A. Escobar (1998). *Statistical Methods for Reliability Data*. John Wiley & Sons, Inc.

Meeker, W. Q., Hahn, G.J., and L. A. Escobar (2017) *Statistical Intervals: A Guide for Practitioners and Researchers*, Wiley.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B* **56**, 3–48.With discussion and a reply by the authors.

Olwell, D. and A. Sorell (2001). Warranty calculations for missiles with only current-status data, using Bayesian methods. In *Proceedings of the Annual Reliability and Maintainability Symposium*, 133–138. IEEE.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics* **9,** 130–134.

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap*. Springer.

Xu, Z., Hong, Y., Meeker, W. Q. (2015), Assessing Risk of a Serious Failure Mode Based on Limited Field Data. *IEEE Transactions on Reliability* **64**, 51–62.