

A Comparison of Maximum Likelihood and Median Rank Regression for Weibull Estimation

Ulrike Genschel
William Q. Meeker
Department of Statistics
Iowa State University
Ames, IA 50011

Abstract

The Weibull distribution is frequently used in reliability applications. Many different methods of estimating the parameters and important functions of the parameters (e.g. quantiles and failure probabilities) have been suggested. Maximum likelihood and median rank regression methods are most commonly used today. Largely because of conflicting results from different studies that have been conducted to investigate the properties of these estimators, there are sharp differences of opinion on which method should be used. The purpose of this paper is to report on the results of our simulation study, to provide insight into the differences between the competing methods, and to resolve the differences among the previous studies.

Key Words: Censored data, Least squares, ML, MRR, Reliability

1 Introduction

1.1 Motivation and purpose

The Weibull distribution, described in Section 2, is perhaps the most widely used distribution for reliability analysis. In the earlier days of reliability data analysis, before desktop computers and reliability analysis software became available, engineers and statisticians commonly used probability plots to analyze censored life data. A nonparametric estimate of the fraction failing as a function of time, consisting of a point for each failure time, would be plotted on specially prepared papers designed such that a Weibull distribution would be represented by a straight line on the plot. If the plotted points did not deviate too much from a straight line, one would draw a line through the points to estimate the Weibull distribution. Nonparametric estimators used for this purpose include (but are not limited to)

- Kaplan and Meier (1958)
- Herd (1960), described more completely in Johnson (1964), and
- Nelson (1969)

As described by Nelson (1982, page 118), except when estimating lower tail probabilities, there is little difference among these methods and in all cases such differences are not important relative to variability in the data-generating process.

After computers became available, it was possible to fit the line on the probability plot by using an objective analytical method. Two approaches emerged. Most statisticians used or advocated the use of maximum likelihood (ML) because of its well-known distributional optimality properties in “large samples.” Many engineers, however, used ordinary least squares (OLS) to draw the line on the probability plot because it was easier to program and more familiar (being covered in most introductory statistics text books). Also, OLS produces a visually appealing line through the points, when the points fall on a line. The ML estimate also provides a visually appealing line through points except in cases where it should not, as will be illustrated in our examples. Intuitively, using OLS to fit the line would not seem to be a good idea because this application violates the assumptions under which OLS is usually justified (constant variance and independent observations) as statistically optimum for estimation. Moreover, OLS regression estimators are linear estimators that put large weight on the extreme observations having large variance. The most commonly used implementation of this OLS/probability plot approach uses median-rank plotting positions (given by Herd 1960 and Johnson 1964) and the method is known today as median-rank regression (MRR).

In July 2000 WQM was asked to respond to a letter from R. B. Abernethy, sent to G. J. Hahn, commenting on Doganaksoy, Hahn and Meeker (2000) and questioning the use of ML estimation in product life analysis. In that same year, WQM received a request from a client at a manufacturing company to provide feedback on a memo. The memo argued strongly for the use of MRR estimation and against the use of ML estimation and had been distributed widely to engineers and statisticians within the company. The arguments were similar to those in Abernethy (1996). The memo concluded:

My recommendation is that we use MRR for populations less than 500 with fewer than 100 failures; and to use MLE for populations of 500 or more, having 100 or more failures.

These events caused us to think more deeply about the various differences of opinion. Although it is possible to find or construct alternative estimators that are better than ML for particular situations, our experience has been that in most relatively simple situations with a fixed number of parameters, it is hard to beat ML. For samples of moderate size (e.g., 20 or 30), it seems to be impossible to find anything that will be consistently better than an ML estimator. Our thoughts were, however, that censoring or extrapolation or special properties of the Weibull distribution could have led to misleading statistical intuition. Thus, over the following year we designed and conducted an extensive simulation experiment to study and compare the properties of ML and MRR estimators. Our simulation results showed a strong preference for the ML method for situations arising in practical reliability analysis. A summary of these results was later reported by Genschel and Meeker (2007).

Meanwhile, we have had other experiences where ML has been questioned and MRR suggested as the best alternative. For example, many statisticians were surprised when MINITAB changed the default estimation method for reliability analysis from ML to MRR (starting with their release 14 and continuing with release 15).

The purpose of this paper is not only to report on the results of our simulation study, but also to provide some insight into the differences between ML and MRR (particularly in small samples) and to resolve differences among other existing studies.

1.2 Limitations of point estimates and point predictions

Although the focus of this paper is to study the properties of point estimators of Weibull distribution parameters and functions of parameters, we want to emphasize that point estimates, by themselves, have limited usefulness. In almost all practical applications, but particularly in reliability applications where there can be safety issues and where only small-to-moderate sample sizes are available, it is essential to quantify uncertainty. Well developed methods for quantifying statistical uncertainty (i.e. uncertainty due to limited data) are available and widely used. Quantifying other kinds of uncertainty (e.g., model error) are more difficult, but still important. Knowledge of the statistical uncertainty provides, at least, a lower bound on the overall uncertainty. See page 5 of Hahn and Meeker (1991) for further discussion of this point.

One reason to study the properties of point estimators is that some statistical interval procedures are based on a specific point estimation method. When a method has good statistical properties, one might expect that the associated interval method would also have good properties and vice versa. Of course one could and probably should study the properties of the interval method directly, as in Jeng and Meeker (2000).

1.3 Previous work on Weibull estimation

During the 1960s and 1970s there were many publications, in addition to those mentioned above, describing research on estimating the parameters of the Weibull distribution. Much

of this early work is summarized in Mann, Schafer, and Singpurwalla (1974). The *Weibull Analysis Handbook* by Abernethy, Breneman, Medlin, and Reinman (1983) describes practical tools, methods, and applications for using the Weibull distribution to analyze reliability data and to make decisions based on the analyses. Subsequent versions of this material, with additions and subtractions, have been published as *The New Weibull Handbook* with the latest edition being Abernethy (2006).

Since 1982, numerous books have been written on statistical methods for reliability data analysis and most of these include some combination of statistical theory, methodology, and applications for using the Weibull and other distributions in reliability applications. There are too many of these to mention all of them, but some of the most important of these include Lawless (1982, with a new edition in 2003), Nelson (1982, with a new updated paperback edition in 2004), Crowder, Kimber, Smith, and Sweeting (1991), Tobias and Trindade (1995), and Meeker and Escobar (1998).

A number of studies have been conducted to compare different methods of estimating the Weibull parameters and functions of these parameters. We will defer discussion of these until after we have presented and explained the results of our study.

1.4 Overview

The remainder of this paper is organized as follows. Section 2 introduces the Weibull distribution and describes different kinds of censoring that can arise in reliability data analysis. Section 3 discusses linear estimation of Weibull distribution parameters and median rank regression. Section 4 describes ML estimation. Section 5 gives details of the design of our simulation experiment. Section 6 presents examples of the analysis of simulated data like those in our study and presents the results of a small simulation to compare type 1 and type 2 censoring. Section 7 summarizes the results of our main simulation study. Section 8 describes other studies that have been done to evaluate the properties of Weibull parameter estimates and compares them with the results of our study. In Section 9 we state some conclusions and recommendations from our study and suggest areas for further research.

2 The Weibull Distribution and Censored Data

2.1 The Weibull Distribution

The Weibull cumulative distribution function (cdf) can be expressed as

$$\begin{aligned} F(t; \mu, \sigma) = \Pr(T \leq t; \eta, \beta) &= 1 - \exp \left[- \left(\frac{t}{\eta} \right)^\beta \right] \\ &= \Phi_{\text{sev}} \left[\frac{\log(t) - \log(\eta)}{1/\beta} \right], \quad t > 0 \end{aligned} \tag{1}$$

where $\Phi_{\text{sev}}(z) = 1 - \exp(-\exp(z))$ is the standard Gumbel smallest extreme value (SEV) distribution cdf, $\eta > 0$ is the Weibull scale parameter (approximately the 0.632 quantile), and $\beta > 0$ is the Weibull shape parameter. The second expression shows the relationship between

a Weibull random variable and the logarithm of a Weibull random variable, which follows an SEV distribution. That is $\mu = \log(\eta)$ is the SEV location parameter and $\sigma = 1/\beta$ is the SEV scale parameter.

The Weibull shape parameter, β , tends to be closely related to the failure mode of a product. A value of $\beta < 1$ implies a decreasing hazard function and suggests infant mortality, while $\beta > 1$ implies an increasing hazard function, suggesting wear out. If $\beta = 1$, the hazard function is constant, implying that the conditional probability of failure in a future time interval, given survival to the beginning of that interval, depends only on the size of the interval and not the age of the unit entering the interval.

2.2 Motivation

The Weibull distribution is popular because it provides a useful description for many different kinds of data, especially in engineering applications such as reliability. One physical motivation for the Weibull distribution is that it is one of the limiting distributions of minima. For example, if a system has a large number of components with failure times that will be independent and identically distributed and the system fails when the first component fails, the Weibull distribution can provide a good description of the system's failure-time distribution. If on the other hand there is a dominant failure mode from a single component, then the system's failure time distribution may be better described by some other distribution.

2.3 Censoring types

Censored data (especially right-censored) are ubiquitous in reliability analysis. Reliability data either come from life tests or from field or warranty data. Life test data are almost always from type 1 (time) censored tests because schedule dictates the time at which the test will end. Field data are almost always multiply censored because of competing failure modes, random entry into a study, and variation in use rates. Again, analysis times are usually dictated by schedule. The simulation study described in this paper covers this range of censoring types.

Type 2 censoring arises when a test is terminated after a given number of failures. Such tests, however, are uncommon in practice. Nevertheless, the majority of research on statistical methods for censored data has dealt with type 2 censoring because it is technically simpler. In Section 6.3 we present the results of a small simulation study to demonstrate that it is important to evaluate a statistical procedure under the kind of censoring that will actually be used.

3 Linear Estimation of Reliability Distributions and Median Rank Regression

3.1 BLU and BLI estimators

Mann, Schafer, and Singpurwalla (1974, Chapter 5) discuss best linear unbiased (BLU) and best linear invariant (BLI) estimators. Nelson (1982) focuses his discussion on BLU estimators but points out that the ideas also apply directly to BLI estimators. These estimators are

based on generalized least squares applied to the observed order statistics arising from type 2 censored data (providing optimum estimation, after accounting for non-constant variance and correlation in the order statistics) and are best in the sense of minimizing mean square error when estimating the underlying Gumbel smallest extreme value distribution parameters and quantiles (as mentioned in Section 2.1, the logarithm of a Weibull random variable has a Gumbel smallest extreme value or SEV distribution). These optimality properties do not continue to hold for nonlinear functions of the parameters that are usually of interest, such as Weibull quantiles, but one would expect that good statistical properties carry over. Optimality properties of these estimators are derived under the assumption of type 2 censoring. Such linear estimators can be extended to progressive type 2 censoring, with more than one censoring point, as described in Balakrishnan and Aggarwala (2000).

Computing for these linear estimation methods requires special tables of coefficients. Only limited tables are available in the books mentioned above and, unlike ML, the methods have not been implemented in commonly used commercial software. For these reasons, and the fact that ML estimators have other important advantages, optimum linear estimation is rarely used in practice today.

3.2 Median rank regression and other estimators based on ordinary least squares

MRR is a procedure for estimating the Weibull parameters $\mu = \log(\eta)$ and $\sigma = 1/\beta$ by fitting a least squares regression line through the points on a probability plot. The analytical motivation for MRR is that the log of the Weibull p quantile is a linear function of $\Phi_{\text{sev}}^{-1}(p)$. That is,

$$\log(t_p) = \mu + \Phi_{\text{sev}}^{-1}(p)\sigma$$

where $\Phi_{\text{sev}}^{-1}(p) = \log[-\log(1-p)]$. MRR estimates are computed by using OLS where the response is the log of the r failure times and the explanatory variable is $\Phi_{\text{sev}}^{-1}(p_i)$, with $p_i, i = 1, \dots, r$ corresponding to the median rank plotting positions (estimates of the fraction failing at the i th ordered failure time).

MRR, BLU, and BLI estimators can all be expressed as linear functions of the log failure times. Unlike BLU and BLI estimators, weights in the MRR linear functions are based on the incorrect assumption of uncorrelated, equal variance residuals. Thus one would expect that MRR estimators have inferior properties, particularly with respect to variability.

There are several different ways to compute the median rank plotting positions. We follow the approach outlined on pages 2-7 of Abernethy (2006). We start by ordering all times in the data set (failures and censoring times) from smallest to largest. Ranks of these ordered times are denoted by i and range from 1 to n . Reverse ranks corresponding only to the failure times are given by $R_k = n - i_k + 1, k = 1, \dots, r$, where i_k is the rank (among all n times) of the k th failure. The k th *adjusted* rank (adjusting for the censored observations, if any), corresponding to the k th failure, can be computed from the recursive formula

$$R_k^A = \frac{R_i \times R_{k-1}^A + n + 1}{R_k + 1}, \quad k = 1, \dots, r.$$

The MRR estimation in our simulation employed the commonly-used approximation to the median rank plotting positions

$$p_k = \frac{R_k^A - 0.3}{n + 0.4}, \quad k = 1, \dots, r$$

due to Benard and Bosi-Levenbach (1953).

3.3 Motivation for MRR

As mentioned in Section 1.1, the original motivation for using OLS (including MRR) to estimate Weibull parameters was simplicity and ease of programming. Today these are no longer valid reasons.

Another argument put forward against ML estimation and in favor of MRR is that ML estimators are biased and behave anti-conservatively (i.e. give optimistic estimates). MRR estimators are also biased, and not always conservatively. In almost all cases the bias in these estimators is dominated by the variance and when the overall accuracy of an estimator is evaluated, under any reasonable criterion for comparing estimators, ML methods are better in almost all practical situations.

Note that averages and bias properties of estimators are misleading if variability is neglected. This is the message behind the joke involving the statistician who, with his head in the oven and feet in the freezer, claims to feel fine on average. Those who argue for estimators with smaller bias, without considering variability properties, make an analogously incorrect argument.

A third reason given for using MRR is that the fitted line goes through the plotted points. The ML estimate line also goes through the points when the Weibull model is appropriate. The MRR estimate will go through the points even when it should not, giving highly misleading results.

When WQM was teaching a short course at the 1999 Fall Technical Conference, a student concerned about the use of ML and expressing a preference for MRR provided an example where the MRR estimate went through the points but the ML estimate did not. The data consisted of three failures and eight censored observations at 1100 hours. Figure 1 is a Weibull probability plot of the data showing both ML and MRR estimates. It is interesting to note that the ML line crosses $0.279 \approx 3/11 = 0.273$, the approximate fraction failing at 1100 hours! This approximation is a general property of ML estimators from type 1 censored life tests. The ML estimate suggests that the Weibull model is inappropriate for these data. The MRR estimate, however, goes through the points and gives seriously incorrect results.

The real story here, previously unknown to the owner of the data, is that the early failures were caused by defective bearings and the test ended before the real life-limiting failure mode was seen. One of the many problems with MRR is that it completely ignores the important information contained in the *position* of the censored observations that occur after the last failure.

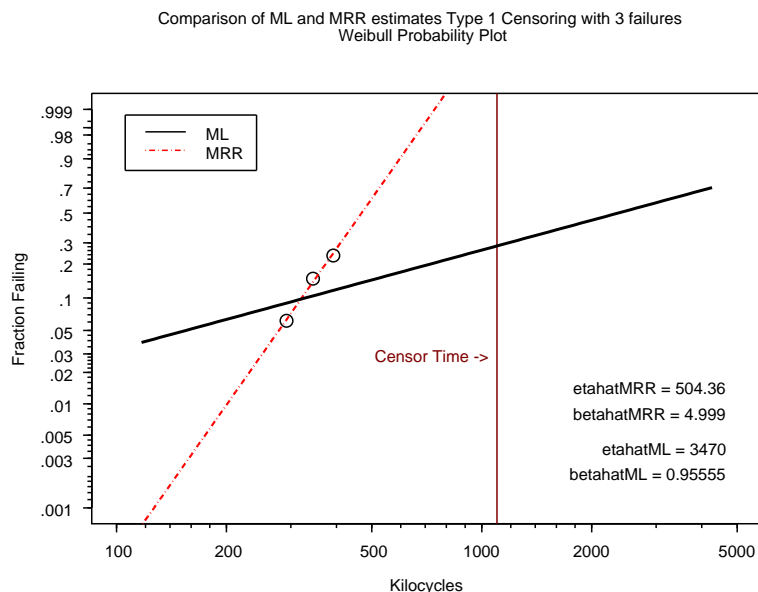


Figure 1: Comparison of MRR and ML Weibull estimates for the Bearing A life test data on Weibull probability paper.

4 Maximum Likelihood (ML) Estimation

4.1 Computing ML estimates

Methods for computing ML estimators and various examples can be found in any of the text books mentioned in Section 1.3. We will not repeat this information here.

4.2 Motivation for ML

With modern computer technology, ML has become the workhorse of statistical estimation. Estimation methods (particularly the method of moments and OLS) taught in introductory statistics courses cannot or should not be used with complicated data (e.g., censored data). Interestingly, many of the common estimators that we present in introductory courses, such as the sample mean to estimate the population mean and OLS to estimate the coefficients of regression model (both linear or nonlinear in the parameters) under the assumption of constant-variance, independent, normal residuals, are also ML estimators! The reasons that ML estimators are so important and so widely used are:

- Under mild conditions, met in most common problems, ML estimators have optimum properties in large samples. Experience, including many simulation studies, has shown that ML estimators are generally hard to beat consistently, even in small samples.
- ML is versatile and can be applied when complicating issues arise such as interval censoring (even with over lapping intervals) or truncation, which arises when only limited information is available about units put into service in the past.

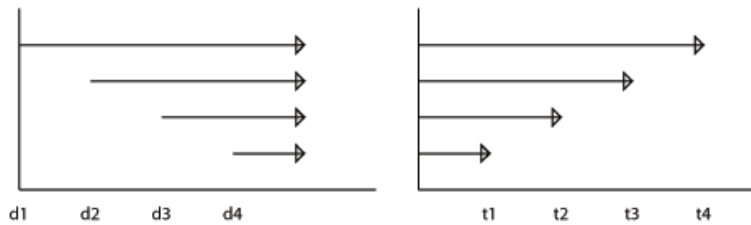


Figure 2: Relationship between staggered entry on the left with analysis at a given time (in real time) and superimposed type 1 censored life tests with different censoring times on the right (in operating time).

- The theory behind ML estimation provides several alternative methods for computing confidence intervals. These range from the computationally easy Wald method to the computationally intensive likelihood-based intervals and parametric bootstrap intervals, based on (approximate) pivotal quantities. Confidence interval procedures based on better estimation procedures will lead to better confidence interval procedures (i.e., shorter expected length for a given coverage probability). Exact confidence interval procedures are available for type 2 censoring.
- ML methods can be used to fit regression models often used in accelerated testing (e.g., Nelson 1990) or to do covariate adjustment for reliability field data.
- With modern computing hardware and software, ML is fast and easy to implement.

For an application of ML involving regression, censoring, and truncation used in a prediction problem, see Hong, Meeker, and McCalley (2009).

5 Design of the Simulation Experiment

5.1 Goals of the simulation

This section provides an explicit description of the design and evaluation criteria for our simulation experiment to compare MRR and ML estimation methods. We designed and conducted our simulation experiment to study the effect of several factors on the properties of estimators for the Weibull parameters and for various quantiles of the Weibull distribution.

5.2 Factors

Our simulation was designed to mimic insertion of a given amount of a product into the field at m equally-spaced points in time (staggered entry), and where the data are to be analyzed at a pre-specified point in time. As shown in Figure 2, this experiment can also be viewed as the superposition of m separate life tests using type 1 censoring with different censoring times.

The particular factors used were

- m : the number of censoring (or product insertion) times,
- $E(r)$: the nominal expected number of failures before time t_{c_m} ,
- $F(t_{c_m})$: the probability that a unit starting at time 0 would fail before reaching the largest censoring time t_{c_m} ,
- $\beta = 1/\sigma$: the Weibull shape parameter.

Without loss of generality (because it is only a scale factor) we used $\eta = 1$ in our simulation.

In Section 7, where we present the results of our main simulation, we will display the important properties of the different estimators as a function of $E(r)$. We do this because the amount of information in a censored sample and the convergence to large-sample properties of estimators is closely related to the (expected) number of failures. If $F(t_{c_m})$ and n were to be used as the experimental factors, there would be a strong interaction between them, making the results much more difficult to present and complicated to interpret. The convergence behavior with respect to n depends strongly on the amount of censoring. Thus the sample size n is not an explicit factor in our simulation, but is a function of $E(r)$, $F(t_{c_m})$, m , and β .

5.3 Factor levels and the generation of censoring schemes

In order to cover the ranges encountered in most practical applications of Weibull analysis, we conducted simulations at all combinations of the following levels of the factors.

- $F(t_{c_m}) = 0.01, 0.05, 0.1, 0.25, 0.40, 0.5, 0.6, 0.75, 0.9, 1.0$
- $E(r) = 4, 5, 6, 7, 10, 15, 20, 25, 50, 100$
- $m = 1, 3, 12, 24, 36$
- $\beta = 1/\sigma = 0.8, 1.0, 1.5, 3.0$

When $m = 1$, we have the special case of type 1 censoring while the case $F(t_{c_m}) = 1.0$ corresponds to no censoring and $t_{c_m} = +\infty$.

For each combination of the factor levels above we simulated and computed both ML and MRR estimates for each of 10,000 data sets. These results were saved in files for subsequent exploration and summarization.

For a given m , $F(t_{c_m})$, and nominal $E(r)$, we can simulate the staggered entry by defining a particular ‘‘censoring scheme’’ consisting of the m superimposed type 1 censored tests with specified allocations and censoring times. This is illustrated in Figure 2 for $m = 4$. Here t_{c_1}, \dots, t_{c_4} are the censoring times on the operating-time scale and $\tau_{I_1}, \dots, \tau_{I_4}$ are the product-insertion times in the real-time scale. The left hand side of Figure 2 illustrates the real-time staggered entry process we are mimicking. On the right is the equivalent superposition of four type 1 censored life tests.

For a given $F(t_{c_m}) < 1$, m equally-spaced censoring times were chosen between 0 and t_m where $t_m = \exp(\Phi_{\text{sev}}^{-1}(F(t_{c_m}))(1/\beta))$, the $F(t_{c_m})$ quantile of a Weibull distribution with

parameters $\eta = 1$ and β . Units are allocated uniformly to the m tests, at the maximum level, such that the actual $E(r)$ is less than the nominal $E(r)$. Then, starting with the shortest test, one additional unit is added to each test until the actual $E(r)$ value exceeds the nominal $E(r)$. Finally, the last added unit is removed so that the actual $E(r)$ will be less than the nominal $E(r)$. Let n_i , $i = 1, \dots, m$, denote the sample sizes for the m groups (life tests or product insertion times) illustrated in Figure 2. Due to the integer constraint on the n_i values, in the final censoring scheme, the actual $E(r)$ (often non-integer) is not always equal to the nominal $E(r)$. But the values must always be within one of each other. When we plot properties of the estimators, we always plot against the actual $E(r)$ values.

5.4 Random number generation and simulating censored data, estimation, and presentation

A faulty random number generator can lead to incorrect simulation results. Potential problems with random number generators could include a period that is too short or other periodicities and autocorrelation. It is important to know the properties of such generators.

The uniform random number generator used in our study is the portable FORTRAN function `rand` (and its interface), available from <http://www.netlib.org>. This function is based on an algorithm due to Bays and Durham (1976) that uses a shuffling scheme to assure an extremely long period. For our implementation (using shuffling among 32 parallel streams of random numbers), the approximation given in Bays and Durham (1976) suggests that the period should be on the order of 10^{28} .

For the efficient simulation of censored samples from a Weibull distribution, we use the simple algorithm described in Section 4.13.3 of Meeker and Escobar (1998). Estimation was performed using ML and MRR algorithms in SPLIDA (Meeker and Escobar 2004). The MRR algorithm was checked against examples in Abernethy (1996). The ML algorithms have been checked against JMP, MINITAB, and SAS. Exploration and graphical presentation of the results were done in S-PLUS.

5.5 Estimability issues and conditioning

With type 2 (failure) censoring, the number of failures in an experiment is fixed. As mentioned earlier, for most practical applications this is unrealistic. With type 1 (time) censoring, there is always a positive probability of zero failures, in which case neither ML nor MRR estimates exist. ML estimation requires one failure and MRR estimation requires two. We thus discarded any samples in which the number of failures was less than two, making our evaluation conditional on having at least two failures. Table 3 in Jeng and Meeker gives the number of observed samples in their simulation in which there were only 0 or 1 failures. Our results files contain similar information but these counts are not reported here, due to space constraints. These counts are less important here because we are primarily comparing two different estimation procedures and the conditioning is the same for both procedures. The probability $\Pr(r < 2)$ could also be computed or approximated without much difficulty. For $m = 1$ it is a simple binomial distribution probability. For $m > 1$ the relevant distribution is the sum of m independent but non-identically distributed binomial random variables. For $E(r) \geq 10$, $\Pr(r < 2) \approx 0$.

5.6 Comparison criteria

Previous simulation studies to compare Weibull estimation methods have focused on the properties of estimators of parameters and interesting functions of parameters, especially distribution quantiles (also known as B-life values). We have used the results of our simulation study to evaluate the properties of the Weibull shape parameter ($\beta = 1/\sigma$), the SEV scale parameter σ , and various Weibull quantiles ranging between 0.0001 and 0.90.

Proper evaluation of the accuracy of an estimator requires a metric that considers both bias and precision. Bias is important, primarily, as a component of estimation accuracy. The other component is precision, often measured by the standard deviation (SD) of an estimator. It is useful to look at bias to learn how much it affects performance. Discovering that bias is large relative to the standard deviation might suggest that reducing bias could improve overall accuracy.

The most commonly used metric for evaluating the accuracy of an estimator is the mean square error (MSE). The MSE of an estimator $\hat{\theta}$ is

$$\text{MSE} = \text{E}[(\hat{\theta} - \theta)^2] = [\text{SD}(\hat{\theta})]^2 + [\text{Bias}(\hat{\theta})]^2$$

where $\text{Bias}(\hat{\theta}) = \text{E}(\hat{\theta} - \theta)$. It may be preferable to report the root mean square error (RMSE), which has the same units as θ . When comparing two estimators θ_1 and θ_2 it is useful to compute relative efficiency, defined here as $\text{RE} = \text{MSE}(\hat{\theta}_1)/\text{MSE}(\hat{\theta}_2)$. $\text{RE} = 0.70$ implies, for example, that the necessary sample size for a procedure using θ_1 is 70% of that needed for θ_2 to achieve approximately equal overall accuracy.

Evaluation criteria are not limited to MSE and can be defined in other ways. For example

$$\text{LOSS} = \text{E}(|\hat{\theta} - \theta|^p). \quad (2)$$

If p is chosen to be 1, the LOSS is known as mean absolute deviation (MAD). The MAD is sometimes preferred because it is less sensitive to extreme observations. Values of p greater than one tend to penalize more strongly larger deviations from the truth. The reason $p = 2$ is so popular is that it leads to mathematical simplifications, but it is also thought of as a convenient compromise between $p = 1$ and larger values of p . One can also define a relative efficiency as the ratio of LOSS for two competing estimators.

The range of the observed estimates of quantiles can vary over many orders of magnitude (particularly MRR estimates) and the empirical sampling distributions are badly skewed. Replacing the expectation in (2) with a median provides a loss function that is more robust to large outliers or badly skewed distributions. This estimator is discussed in Hampel (1974). With $p = 1$, this is known as the *median absolute deviation* and we refer to this statistic as MdAD. Another alternative for comparing estimators of quantiles with badly skewed sampling distributions is to compute measures of location and spread for the smallest extreme value quantiles (i.e., on the log time scale).

In our evaluations and comparisons, we experimented with all of these alternatives. On the time scale, mean statistics like loss, even with $p = 1$, were unstable because of extreme outliers, especially with MRR estimation. Thus on this scale we used MdAD for evaluation and comparison. Most of our evaluations were done on the log scale using means and MSE as

metrics, as these are easier to interpret. Our overall conclusions do not, however, depend on these choices.

6 Examples of Simulation Details and Comparison of Type 1 and Type 2 Censoring

This section has the dual purpose of illustrating some particular examples of the analysis of simulated data and to give a sense of the differences between ML and MRR estimates. We also present a small side-simulation study focused on showing that the censoring scheme under which estimators are compared (e.g., type 1 versus type 2) can have an effect on the comparison.

The small simulation in this section is based on an example using $n = 60$ specimens to estimate the life of an adhesive in a high-temperature accelerated life test. The assumed Weibull distribution parameters used in this simulation were $\eta = 211.7$ hours and $\beta = 3$ and from equation (1), the probability of failing before the censoring time of $t_{cm} = 100$ hours is $\Pr(T \leq 100) = 0.1$. Thus the expected number of failures in the type 1 censored life test is $E(r) = 0.10 \times 60 = 6$.

6.1 Examples of simulation detail

The simulated data sets analyzed in Figures 3, 4, 5, and 6 were selected from a set of 2000 simulated type 1 censored data sets that will be presented, in summary form, in the next section. These three examples were chosen from the larger set specifically to illustrate what happens when the plotted points (i.e., the nonparametric estimate) do (Figures 3 and 6) and do not (Figures 4 and 5) fall along a straight line.

The thicker, longer line in the plots shows the true Weibull distribution. The thinner solid and dashed lines show the ML and MRR estimates, respectively. The dotted curves are 95% pointwise confidence intervals based on inverting the Weibull likelihood ratio test (e.g., Chapter 8 of Meeker and Escobar). The points are plotted using the approximate median rank positions $(i - 0.3)/(n + 0.4)$. Note that the ML line would have agreed better with the points had the plotting positions $(i - 0.50)/n$ been displayed instead (as in Meeker and Escobar 1998, Lawless 2003, and Somboonsavatdee, Nair, and Sen 2007).

Figure 3 compares ML and MRR Weibull distribution estimates for a simulated type 1 censored sample that resulted in seven failures before the censoring time of 100 time hours. Because the points lie close to a straight line, the ML and MRR estimates agree well. These two estimates, however, deviate importantly from the truth and would give unjustifiably optimistic estimates of small quantiles often needed to make important decisions. If the need were to extrapolate to the right to predict future failures, the predictions would be overly pessimistic. In either case, especially if there are potential safety issues or large losses, it would be vitally important to quantify statistical uncertainty with an appropriate statistical interval. Of course it is important also to recognize that such intervals reflect only statistical uncertainty due to limited data. The actual uncertainty in an estimate is sure to be even larger.

Figure 4 compares ML and MRR estimates for a second simulated type 1 censored sample from the same model and test plan that resulted in six failures before 100 hours. In this

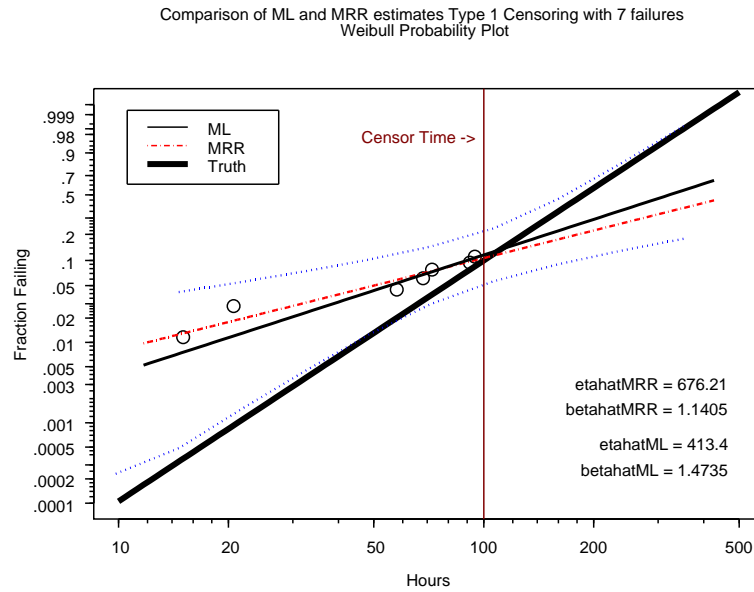


Figure 3: Comparison showing agreement between ML and MRR estimates for a simulated type 1 censored sample that resulted in three failures before $t_{cm} = 100$ hours.

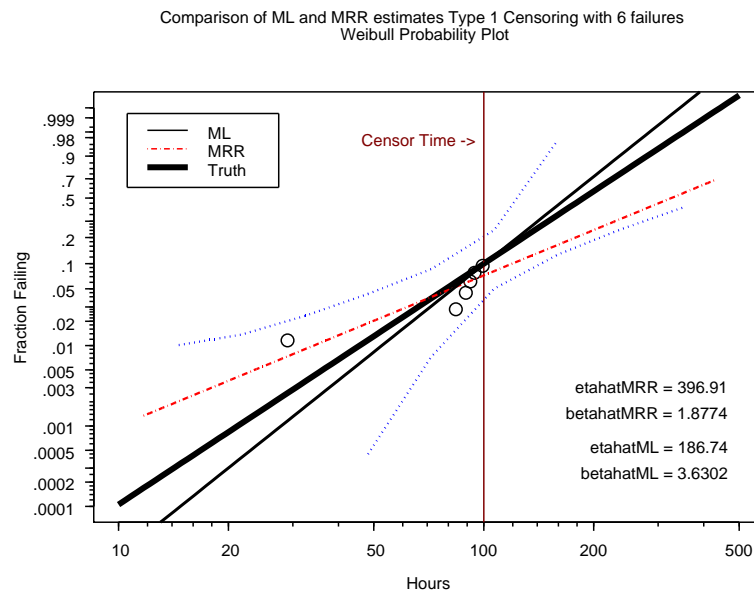


Figure 4: Comparison showing disagreement between ML and MRR estimates for a simulated type 1 censored sample that resulted in six failures before 100 hours.

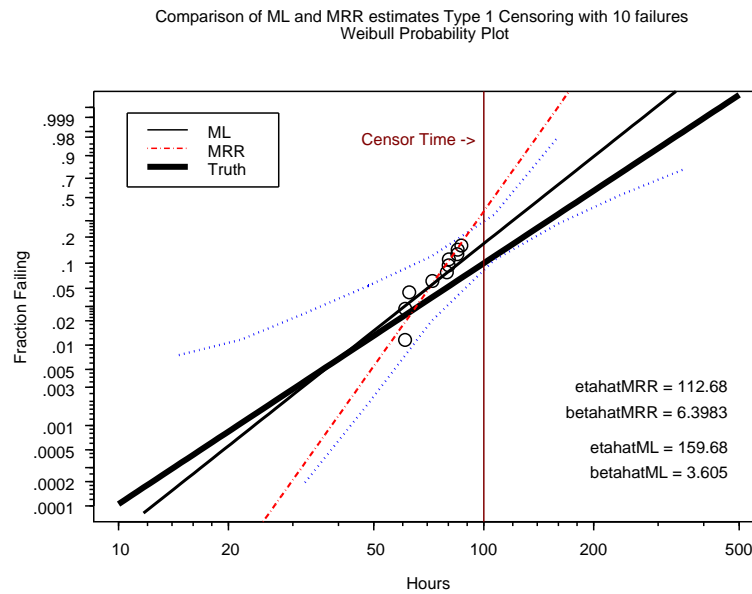


Figure 5: Comparison showing disagreement between ML and MRR estimates for a simulated type 1 censored sample that resulted in ten failures before 100 hours.

case there was an early failure (not surprising given the large variance of the smallest order statistics from a Weibull distribution). As mentioned earlier, MRR gives that observation too much weight when estimating the Weibull distribution parameters. The ML line provides a much better fit.

Figure 5 is similar to Figures 3 and 4 except that in the sample, the first failure came somewhat later than would have been predicted by knowing the true model. Again, MRR gives this observation too much weight and the ML estimate provides a line closer to the truth.

Figure 6 is an extreme example of disagreement that arises where there are two failures near to the type 1 censoring point. The ML estimate is, to a certain extent, tied down because of the constraint that it crosses the censoring point at approximately the fraction failing at that point. The MRR estimator has no such constraint and thus the estimate of β can be extremely large.

Note that, as shown in Hong, Meeker, and Escobar (2008), the likelihood-based confidence intervals in Figures 3, 4, 5, and 6 can be used to obtain a confidence interval on either the fraction failing at a particular point in time (looking vertically) or for a particular quantile (looking horizontally).

6.2 Comparison of ML and MRR under type 1 censoring

Figure 7 is a summary showing ML Weibull distribution estimates for 50 of the 2000 type 1 censored simulations similar to the examples shown in detail in Section 6.1. Again, the longer, thicker line corresponds to the true model ($\eta = 211.7$ hours and $\beta = 3$). Figure 8 is a similar plot showing MRR estimates for the *same* 50 data sets. Comparing these plots carefully, one

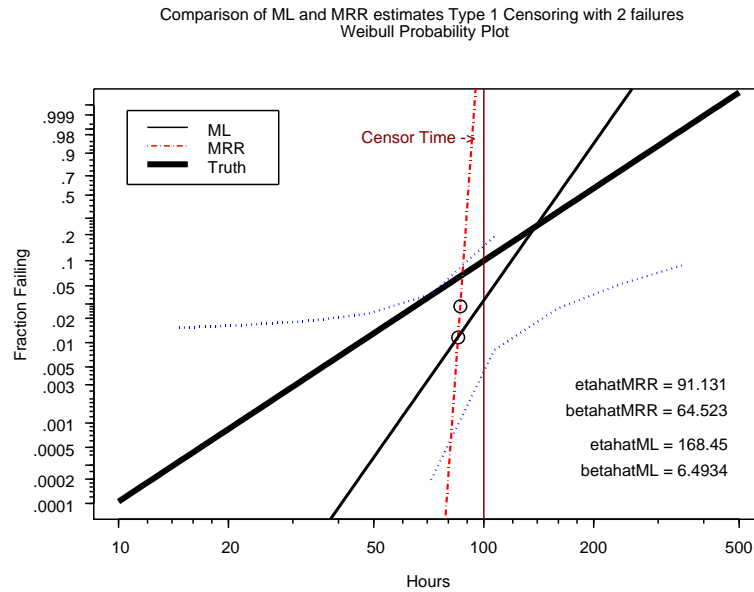


Figure 6: Comparison showing disagreement between ML and MRR estimates for a simulated type 1 censored sample with two late failures.

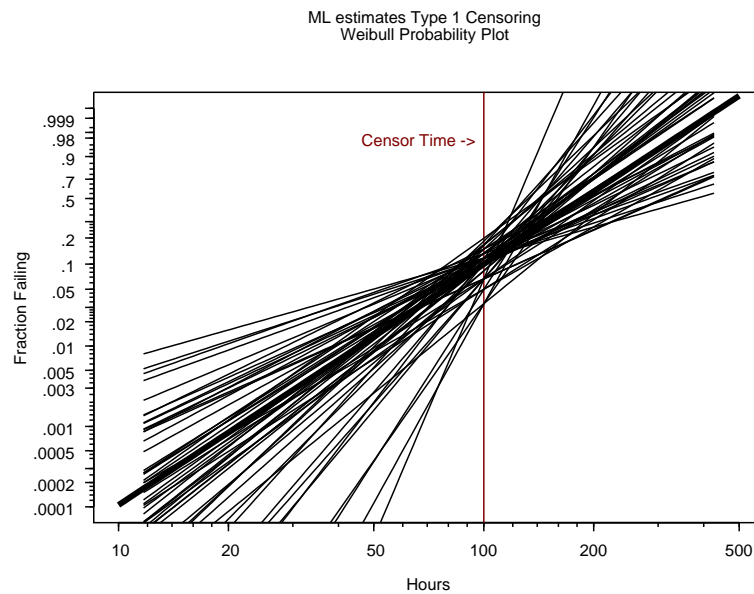


Figure 7: Summary showing 50 ML estimates based on type 1 censored simulated samples.

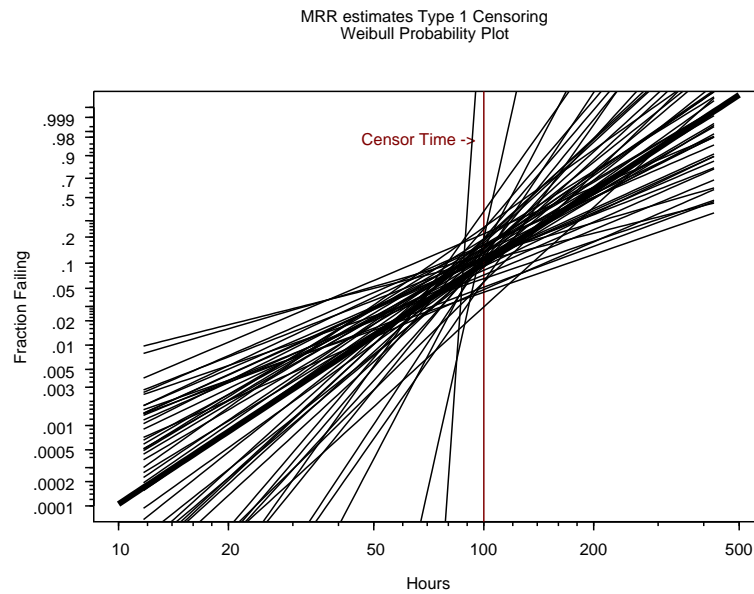


Figure 8: Summary showing 50 MRR estimates based on the same type 1 censored simulated samples used in Figure 7.

can see a higher density (less spread) among the lines near the to the true model (the longer, thicker line) and, generally, more spread in the MRR estimates.

Figure 9 provides a summary of the ML and MRR estimates for all 2000 simulated type 1 censored samples. There are three lines for each estimation method. The center lines (which agree very well with the truth in this case) show the median of estimates for a set of quantiles ranging between 0.0001 and 0.90. Using medians here is favorable toward MRR, relative to other measure of central tendency for skewed distributions that we tried, such as the geometric mean. We saw that any kind of mean-based statistic would be sensitive to outliers in the MRR, giving a stronger indication of bias. The upper and lower curves are the 0.05 and 0.95 quantiles, respectively, of the ML (or MRR) estimates, again for quantiles ranging between 0.0001 and 0.90. This plot shows that, as is often the case, the ML estimator has slightly more median bias than the MRR estimator, but that the MRR estimator has considerably more variability and that variability dominates bias. The non-smooth behavior in the upper MRR line is due to events like that displayed in Figure 6 (which can also be seen in Figure 9).

6.3 Comparison of type 1 and type 2 censoring results

As described in Section 2.3, there are important *practical* differences between type 1 and type 2 censoring (type 1 is commonly used and type 2 is not). In this section we will investigate the statistical differences between these two kinds of tests by comparing the performance of ML and MRR estimators under both type 1 and type 2 censoring. Figure 10 is similar to Figure 9 except that it displays a summary of 2000 ML and MRR estimators under *type 2 censoring*, using the same model as in the other simulations in this section. Regardless of the type of

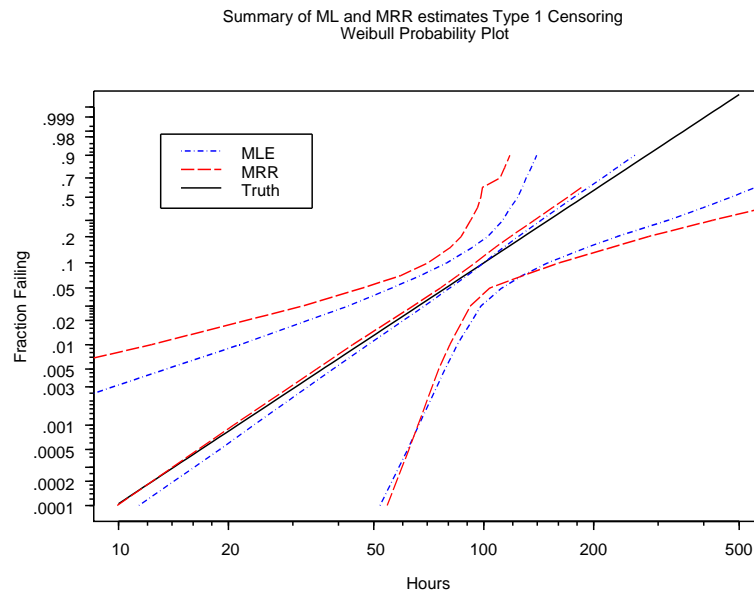


Figure 9: The geometric mean, 0.05 quantile, and the 0.95 quantiles of the type 1 censoring empirical sampling distributions of both ML and MRR estimates of quantiles ranging from 0.0001 to 0.90.

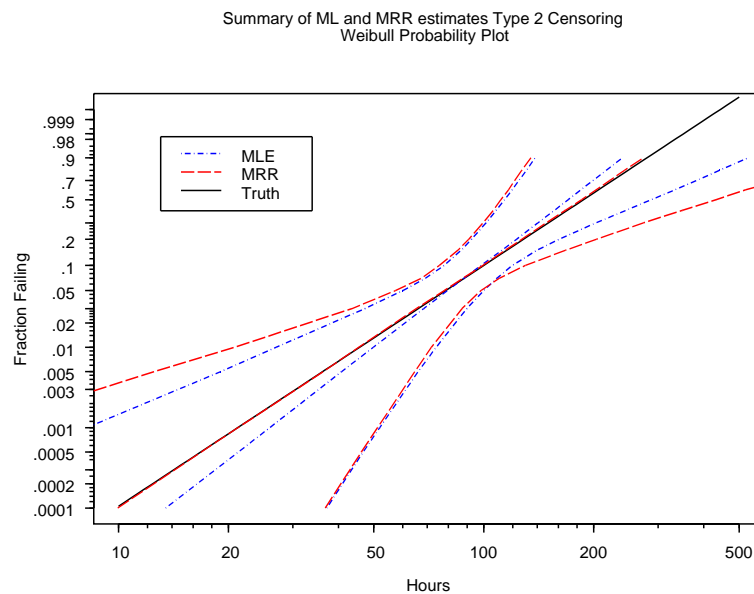


Figure 10: The geometric mean, 0.05 quantile, and the 0.95 quantiles of the type 2 censoring empirical sampling distributions of both ML and MRR estimates of quantiles ranging from 0.0001 to 0.90.

censoring, MRR estimates have more spread when compared with the ML estimates. Careful inspection of these figures shows that there is somewhat more bias in the ML estimates under type 1 censoring for some quantiles, particularly in the upper and lower tail of the distribution. For some combinations of our experimental factors this additional bias can be large enough that the MSE of the ML estimates is larger than that of the MRR estimates for some quantiles.

In our evaluations under type 2 censoring, it is more common to have the MSE of the ML estimates be larger than that of the MRR estimates. One reason for this is that in type 2 censoring, there is no gap between the last failure and the censoring time. For this reason, the probability of extreme events (e.g., Figure 6) is smaller in type 2 censoring when the number of failures is constrained to be four or five. In conclusion, one will get a biased comparison if one evaluates estimators under type 2 censoring if one wants to know the properties under the more commonly used type 1 censoring.

7 Simulation Experiment Results

We generated summary plots to evaluate and compare the sampling properties of the ML and MRR estimates for $\sigma = 1/\beta$, β , and various Weibull quantiles ranging from 0.0001 to 0.90 for all combinations of the experimental factor levels listed in Section 5.3. We made separate sets of summary plots to investigate the effect of using different evaluation metrics (e.g., the usual definition of bias given in Section 5.6 versus median bias and MSE versus MdAD). Although there are differences among these metrics, the overall conclusions remain the same. As described in Section 5.6, we have focused, primarily, on bias and RE for $\sigma = 1/\beta$ and the logarithms of Weibull quantiles. In Section 7.4, however, we present results on the empirical sampling distributions of estimates for β and the quantile on the time scale.

7.1 General observations

We studied an extensive set of evaluation plots for RE for estimating σ and the log Weibull quantiles across the different combinations of the experimental factor levels. As we did this, similarities and patterns emerged that will allow us to easily summarize the results with a small subset of the large number of figures that we produced. In particular the graphics for RE for different values of β were, to the eye, exactly the same. Similarly, plots of the means of these estimates, relative to the true value, as a function of $E(r)$ were, to the eye, almost exactly the same. The reason for this is that RE and relative bias (i.e., bias in the log scale divided by σ) are invariant to changes in β with complete data and type 2 or progressive failure censoring and *approximately* so for our time-censored samples. This can be shown using results in Escobar (2009) based on the equivariance properties of ML and MRR estimators and that RE and relative bias are both functions of pivotal quantities under type 2 censoring (i.e., they do not depend on either η or σ). Similarly, the results for different values of m (the number of superimposed type 1-censored samples) were not substantially different (i.e., the general patterns were the same). Thus we will primarily discuss the cases $m = 1$ (type 1 censoring) and $\beta = 1$. There are important differences relative to the factor $F(t_{c_m})$, the expected fraction failing by t_{c_m} . The patterns across the different levels of this factor are, however, predictable

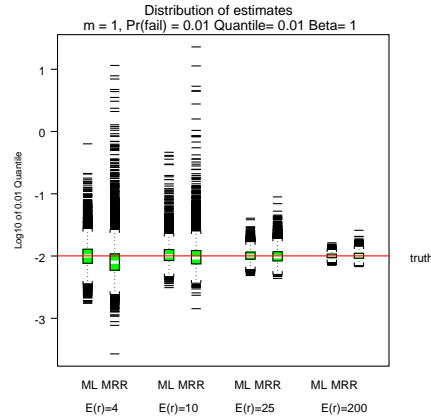


Figure 11: A comparison of ML and MRR sampling distributions of $t_{0.10}$ for different values of $E(r)$ under type 1 censoring for $F(t_{c_m}) = 0.01$, $\beta = 1$, and $m = 1$. The white line indicates the position of the median.

enough that we can summarize the results with using just the plots for the extreme levels in our simulation, $F(t_{c_m}) = .01$ and 1.0.

7.2 Boxplots illustrating selected sampling distributions

Figure 11 displays pairs of box plots to compare the empirical sampling distributions of ML and MRR estimators for the 0.01 quantile for several values of $E(r)$ with $F(t_{c_m}) = 0.01$, $m = 1$, and $\beta = 1$. Figure 12 is similar to Figure 11, providing a summary of ML and MRR estimators for the 0.50 quantile for $F(t_{c_m}) = 1.0$ (i.e., no censoring), $m = 1$, and $\beta = 1$. This appears to be the set of factor-level combinations most favorable to MRR. Of course reliability data sets with no censoring are rare.

Figure 11 shows that the MRR estimates have much more variability than ML estimates. Note that for small $E(r)$, the sampling distributions of the estimates of the quantiles can range over many orders of magnitude and is highly skewed, even on the log scale. Even though the complete-data conditions behind Figure 12 are favorable toward MRR, MRR still does poorly relative to ML. Figure 11 is more typical of other points in our factor space.

7.3 Relative efficiency and bias estimates of σ and log Weibull quantiles

Figure 13, for $F(t_{c_m}) = 1.0$, shows the relative efficiency $RE = \text{MSE}(\hat{\theta}_{ML})/\text{MSE}(\hat{\theta}_{MRR})$ for $\sigma = 1/\beta$ and the 0.10, 0.50, and 0.90 quantiles. For estimating σ , $RE \approx 0.75$ for all values of $E(r)$. The shape and level of the RE relationship was similar for all other levels of $F(t_{c_m})$ (and m and β).

For estimating the quantiles, RE follows an interesting pattern. In particular, when estimating in the lower or the upper tail of the distribution, RE is relatively low. When estimating the 0.50 quantile, however, RE is close to 1.

There is a similar pattern in Figure 14, where $F(t_{c_m}) = .01$. Now the RE is close to 1

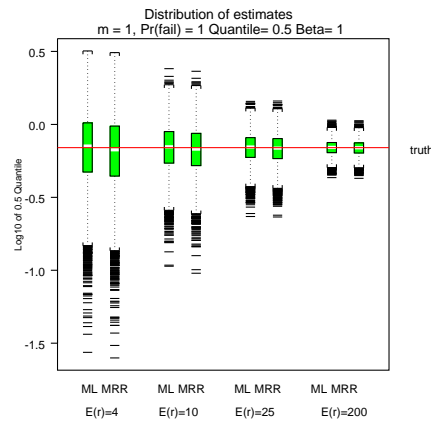


Figure 12: A comparison of ML and MRR sampling distributions of $t_{0.50}$ for different values of $E(r)$ under type 1 censoring for $\beta = 1$ and $m = 1$, and $F(t_{c_m}) = 1.0$. The white line indicates the position of the median.

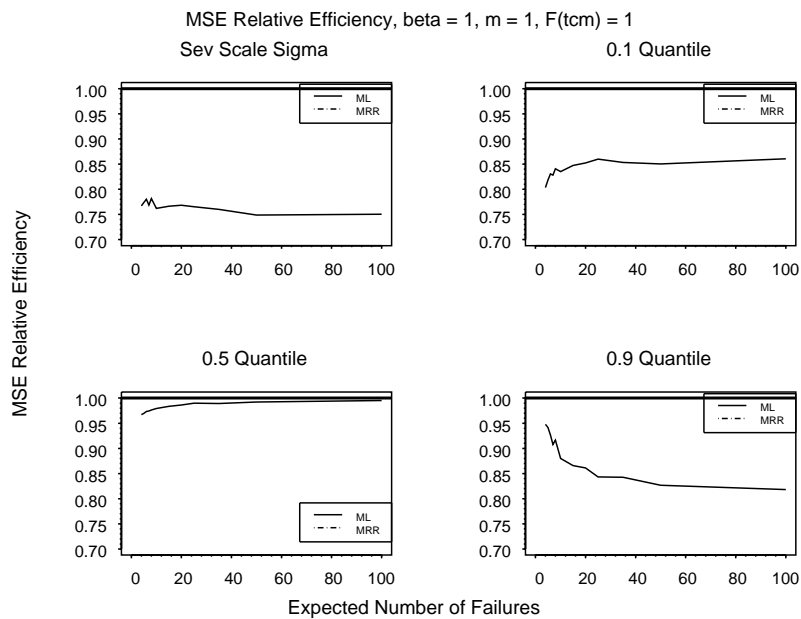


Figure 13: $RE = \text{MSE}(\text{MRR})/\text{MSE}(\text{ML})$ versus $E(r)$ for $F(t_{c_m}) = 1.0, m = 1, \beta = 1$.

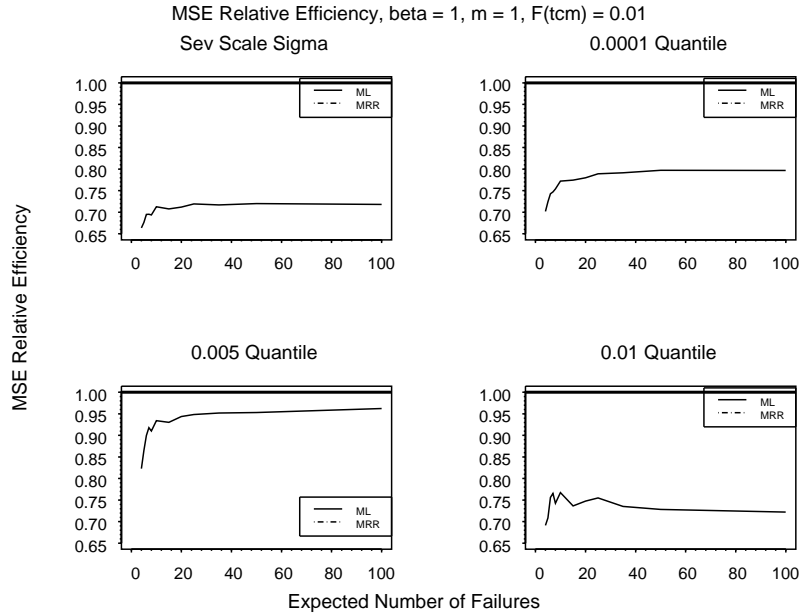


Figure 14: $RE = \text{MSE}(\text{MRR})/\text{MSE}(\text{ML})$ versus $E(r)$ for $F(t_{c_m}) = 0.01$, $m = 1$, $\beta = 1$.

when estimating the 0.005 quantile, but much smaller for the 0.0001 and 0.01 quantiles. For larger quantiles, the RE does not vary much (with the particular quantile or $E(r)$) and typically at a level near 0.75). The pattern is the same for all levels of $F(t_{c_m})$. That is, the RE is at its highest level (approaching but generally not exceeding 1) for quantiles that are close to $F(t_{c_m})/2$. This is in agreement with what we can see from Figures 9 and 10 where (recalling that the censoring time was at 100 hours and the expected fraction failing is 0.10) the spread in the ML and MRR estimates is about the same and the differences in performance between type 1 and type 2 is small around the 0.05 quantile.

Figures 15 and 16 are parallel to Figures 13 and 14, for $F(t_{c_m}) = 1.0$ and 0.01, respectively, the sample mean of the estimates divided by their true values, as a function of $E(r)$.

We have not directly addressed Monte Carlo errors in our results. With a sample size of 10,000, Monte Carlo error will be negligible for mean statistics when the sampling distributions are not too variable. Of course, when $E(r)$ is small, the sampling distributions are sometimes highly variable. Because the evaluations at the different values of $E(r)$ were done independently, the smoothness (or lack thereof) in our plots indicates the degree of noise. Of course, median statistics have more Monte Carlo error than mean statistics, as we will see in the next section. Even in this case, however, the additional error will not substantially cloud our results.

7.4 Relative efficiency and bias estimates of β and Weibull quantiles

Figures 17, 18, 19 and 20 are similar to Figures 13, 14, 15 and 16, except that they provide evaluations for estimates on the time scale. As mentioned earlier, we could not use mean-type evaluations to evaluate the properties of the empirical sampling distributions due to the extremely long tails will in the distributions, as we saw in Figures 11 and 12. These evaluations

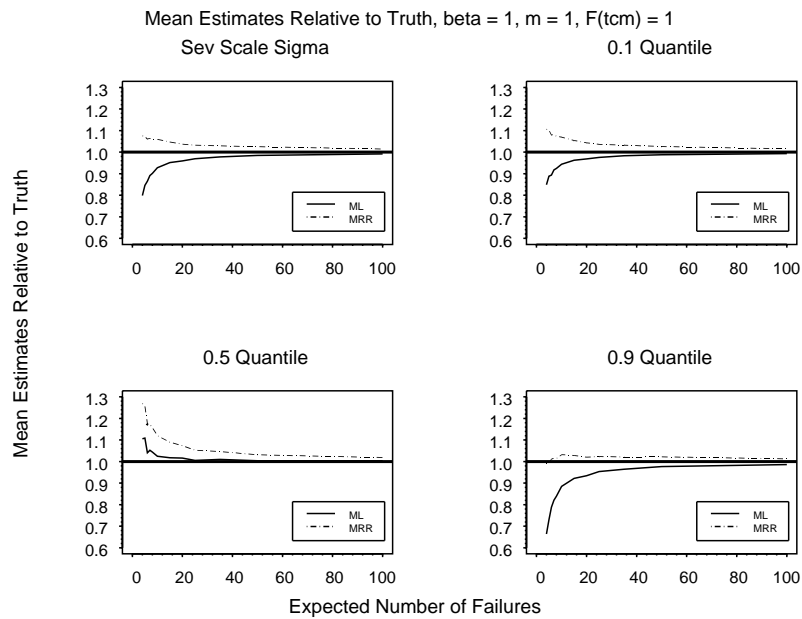


Figure 15: Mean Estimates, relative to the true value versus $E(r)$ for $F(t_{c_m}) = 1.0$, $m = 1$, $\beta = 1$.

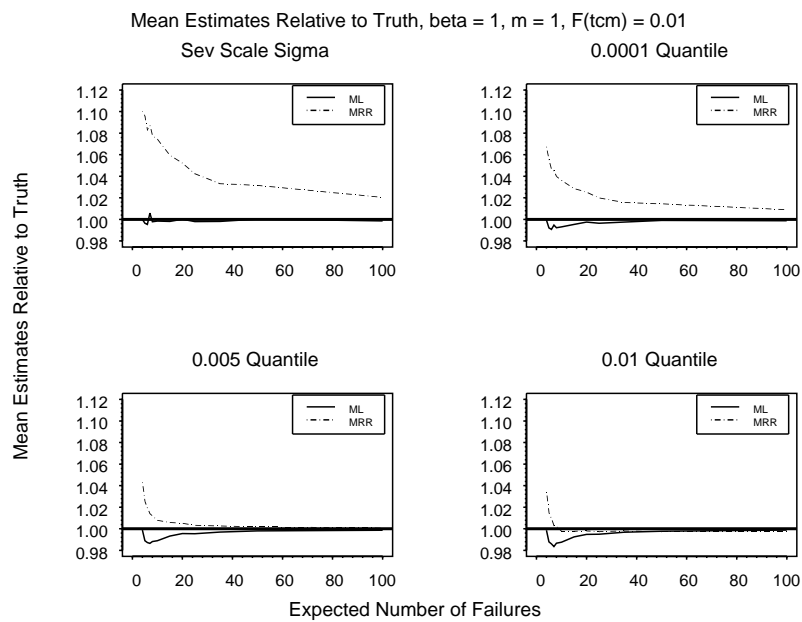


Figure 16: Mean Estimates, relative to the true value versus $E(r)$ for $F(t_{c_m}) = 0.01$, $m = 1$, $\beta = 1$.

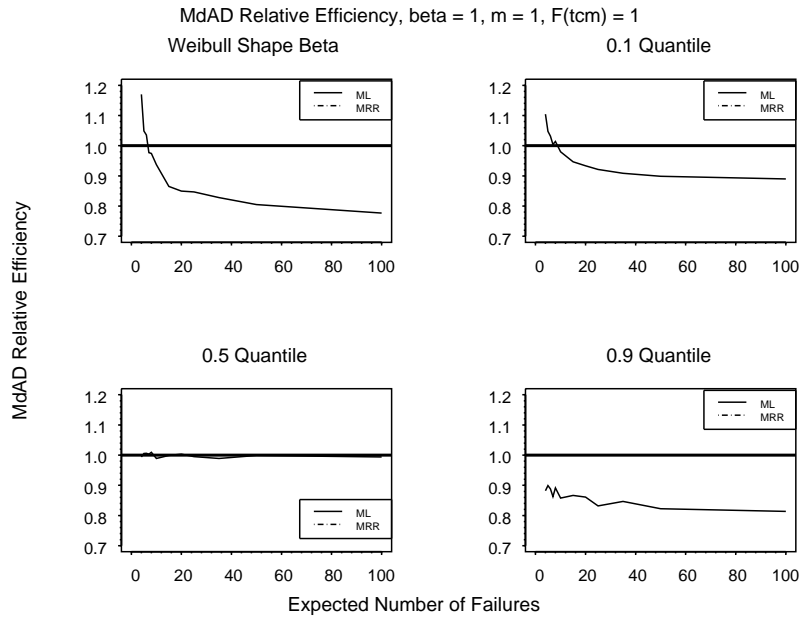


Figure 17: $RE = \text{MdAD}(\text{MRR})/\text{MdAD}(\text{ML})$ versus $E(r)$ for $F(t_{c_m}) = 1.0, m = 1, \beta = 1$.

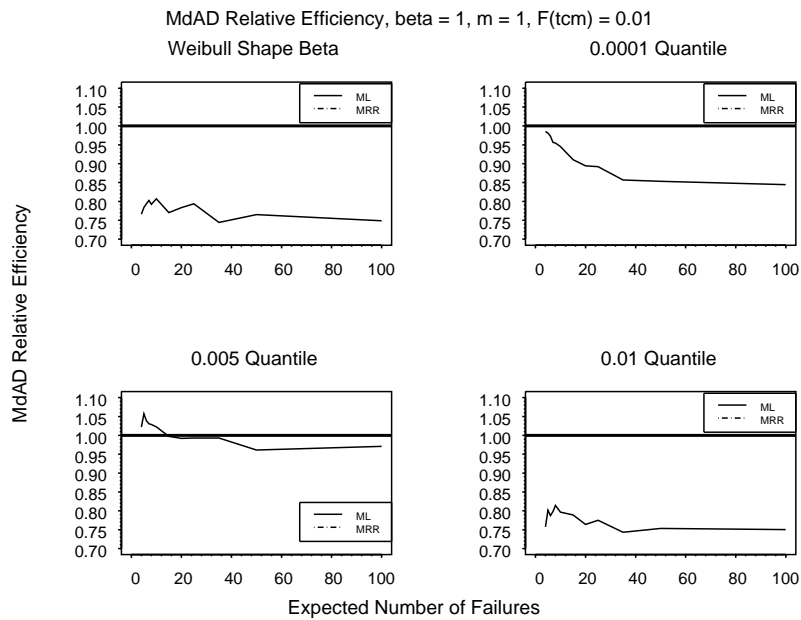


Figure 18: $RE = \text{MdAD}(\text{MRR})/\text{MdAD}(\text{ML})$ versus $E(r)$ for $F(t_{c_m}) = 0.01, m = 1, \beta = 1$.

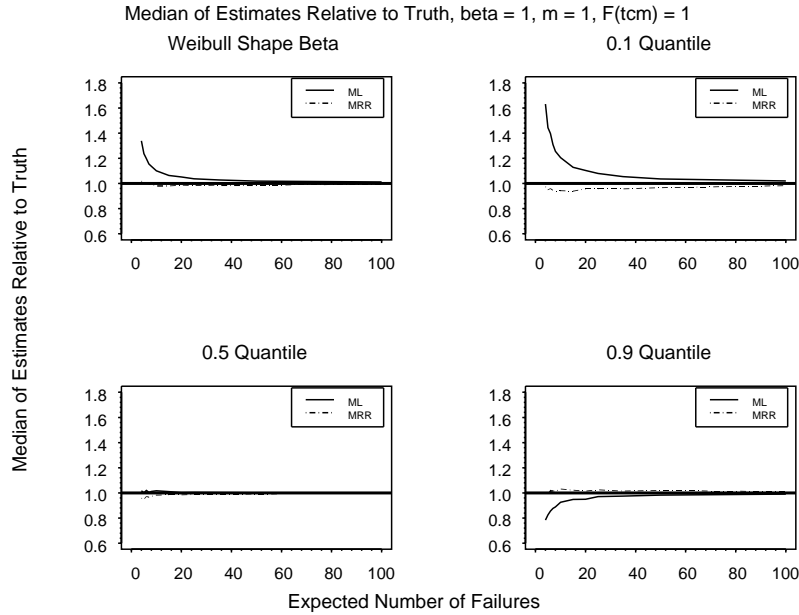


Figure 19: Median estimates, relative to the true value versus $E(r)$ for $F(t_{c_m}) = 1.0$, $m = 1$, $\beta = 1$.

do show that MRR can have a slightly higher RE in some very special cases (e.g., no censoring and estimating quantiles in the center of the distribution), but that overall, the performance of MRR estimators is poor.

8 Reconciliation with Previous Studies

As mentioned earlier, there are sharp differences of opinion whether one should use ML or MRR to estimate Weibull distribution parameters and functions of these parameters. The reason for these differences seem to lie in differences in conclusions from different studies that have been done to compare these estimators. This section reviews some previous studies that have been conducted. Comparisons are not straightforward because

- Each study was conducted to mimic a different situation (e.g., type 1 censoring, type 2 censoring, random censoring, single and multiple censoring).
- There are differences in choices of the levels of experimental factors (e.g., different amounts of censoring and different values of the Weibull shape parameter).
- The studies used different evaluation criteria (e.g., mean bias versus median bias and standard deviation versus root mean square error versus mean absolute deviation).

Table 1 summarizes these differences. Nevertheless it is possible to see some consistency in the results among some of these previous studies.

Table 1: Studies Comparing Weibull Estimators.

| Study | Focus | Evaluation Criteria | β | Sample Size Amount Censored | Censoring Type |
|-------------------------------|---|-----------------------------------|-----------------------|---------------------------------------|--------------------------------------|
| Gibbons and Vance (1981) | $\sigma, \beta, t_{0.10}$ | MSE | 1 | $n = 10, 20$ %Fail 30-100 | Type 2 (failure) |
| Abernethy et al. (1983) | $\beta, t_{0.001}$ | Median Bias Standard Deviation | 0.5, 1.0 3.0, 5.0 | $n = 1000, 2000$ $r = 2$ to 10 | Staggered entry Failure censoring |
| Somboonsavatdee et al. (2007) | μ, σ $\log(t_{0.10})$ to $\log(t_{0.90})$ | Relative MSE | 1 | $n = 25$ to 500 E(%Fail) 25-100 | Random Specified distributions |
| Skinner et al. (2001) | β, η | MSE | 1.2, 1.8, 2.4, 3.0 | $n = 5, 10, 15$ %Fail 20-53.3 | Random: Early, Middle, Late |
| Liu (1997) | $t_{0.000001}$ to $t_{0.01}$ | Median Bias MAD, MSE | 0.5, 1.0, 3.0, 5.0 | $n = 10, 25, 50, 100$ %Fail 30-100 | Unspecified MonteCarloSMITH |
| Genschel and Meeker (2009) | $\sigma = 1/\beta, \beta$ $\log(t_{0.0001})$ to $\log(t_{0.90})$ | Mean Bias Relative MSE | 0.5, 1.0, 3.0, 5.0 | Various E(r) E(%Fail) 1-100 | Staggered entry Type 1 (time) |

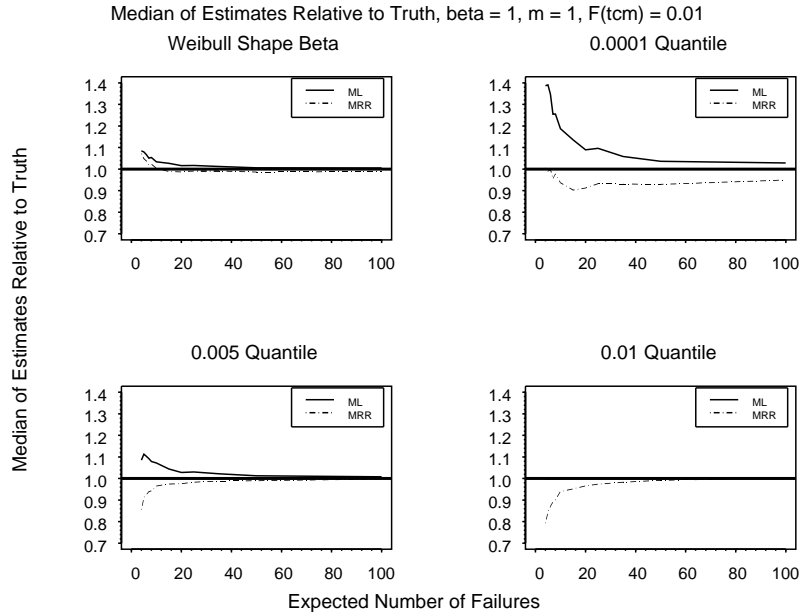


Figure 20: Median estimates, relative to the true value versus $E(r)$ for $F(t_{c_m}) = 0.01$, $m = 1$, $\beta = 1$.

Gibbons and Vance (1981) summarize the results of a large simulation study comparing ML, BLU, BLI, and MRR and several other estimators using type 2 censoring. They report only the results for $\beta = 1$, because the relative comparison was similar for other values of β (as expected from theory). For estimating $\sigma = 1/\beta$ the ML, BLU, BLI estimators have considerably smaller MSE values when compared with the MRR estimators for both $n = 10$ and 25. BLI is slightly better than BLU, as would be predicted from theory and ML is almost the same as BLI. This is not surprising given that BLU and BLI estimators have optimality properties relative to this criterion (but BLU has an unbiasedness constraint) and that ML estimators are highly correlated with BLU and BLI estimators. For estimation of β , in some cases MRR estimators perform better than ML, BLU, and BLI for $n = 10$ but differences are small until the amount of censoring increases to 80% (2 failures). For $n = 25$, there is little difference among the four estimators. For estimation of the 0.10 Weibull quantile, the differences among the estimators is not large, but the MSE for MRR is smaller than that of the others. Given the results of our comparisons between evaluation under type 1 and type 2 censoring in Section 6.3, our simulation results are not inconsistent with those of Gibbons and Vance (1981).

Appendix F of Abernethy, Breneman, Medlin, and Reinman (1983) contains a detailed description of a simulation to compare MRR and ML estimators, mimicking a staggered entry situation similar to the one that we used. The primary difference is that their rule was to analyze the data after a fixed number of failures, rather than a fixed point in time. The fixed number of failures in their experiment ranged between 2 and 10. They observed that both MRR and ML tend to overestimate β (i.e., positive bias), but that MRR always has a larger

standard deviation than ML. For estimation of $t_{0.10}$, MRR tends to underestimate and ML tends to overestimate, but the bias is completely dominated by variance in the MSE. All of these observations are consistent with results from our simulation.

Nair (1984) describes a study to compare asymptotic relative efficiency and small sample properties of OLS estimators similar to MRR and symmetric censoring (same amount in each tail). He used the popular plotting positions $(i - 0.5)/n$. Nair (1984) also reviews earlier theoretical work that studies properties of linear estimators based on a subset of the order statistics (which arise in certain kinds of failure censoring). This work was extended in Somboonsavatdee, Nair, and Sen (2007) who consider random right censoring with specified censoring distributions. Such censoring schemes would mimic censoring arising from random phenomena like competing failure modes, random entry into a study, and variation in use rates. For plotting positions they used a generalization $(i - 0.5)/n$ that uses the point halfway up the jump in the Kaplan-Meier estimate, also suggested in Lawless (2003) and Meeker and Escobar (1998). Their evaluations of quantile estimators are, like ours, computed on the log scale because that is the scale used for pivoting to obtain confidence intervals. For the Weibull distribution their conclusions are that OLS estimators have much lower relative efficiency, both asymptotically and for finite sample sizes, when compared to ML. Interestingly, the only case in their study where the OLS estimators are as good as ML estimators is for the lognormal distribution with no censoring. Our results for the Weibull distribution are consistent with theirs.

Skinner, Keats, and Zimmer (2001) describe a small simulation to compare ML, MRR and another estimator for the Weibull parameters. They generated censored data by randomly choosing binary patterns from a specified set to determine which observations should be censored or not. They used different sets of patterns in order to compare the properties of the estimators with early, middle and late censoring within the sample of observations. Their simulation showed that ML always had smaller MSE than MRR for estimating η . For estimating β , however MRR had smaller MSE values when censoring was concentrated at the beginning or in the middle, but not at the end. The results in this study for estimation of β seem at odds with our study and others. We suspect that this is because of the different method that was employed to generate censored samples.

Liu (1997) conducted an extensive simulation study comparing estimation methods for the Weibull and lognormal distributions for complete and censored data. His results comparing RMSE for ML and MRR are, for the most part, highly favorable toward MRR relative to ML for estimating Weibull parameters and quantiles, even for sample sizes as large as 100. These results are inconsistent with our results and any other simulation results that we have seen. We attempted, without success, to learn how the censored samples for this study were computed. Lui (1997) says only that he used MonteCarloSMITH to do his simulations. If we knew precisely how the censored samples had been generated, we could try to reproduce the results and learn the root cause of the differences.

Another study, by Olteann and Freeman (2009) has also been completed and is to be published in the same issue as this article.

9 Conclusions, Recommendations, and Areas for Further Research

The main conclusions from our study are as follows.

- When evaluated under appropriate criteria (e.g. MSE or some other similar metric that takes variation into consideration and at least approximates the users true loss function), ML estimators are better than MRR estimators in all but a very small part of our extensive evaluation region.
- There are important differences between evaluating an estimation procedure under type 1 and type 2 censoring. ML has an advantage in type 1 censoring in that it uses the information contained in the location of the censored observations. MRR ignores this, as we saw in Figure 1. This information is particularly important when there are few failures and is one of the reasons that ML out performs MRR.
- All previous studies comparing ML and MRR estimators were different in one way or another. Section 6.3 showed that while ML estimators usually have better precision than MRR estimators in type 2 censoring experiments, the differences are smaller than in type 1 simulations. This suggests that in order to make appropriate comparisons, simulations need to be conducted to carefully mimic the testing or reliability data-generating processes that are in use (rather than choosing a censoring scheme that is convenient).

We have the following recommendations

- Statistical theory should be used to guide the choice of inference methods. Even large sample approximations can be useful in this regard.
- In complicated situations where exact analytical results are not available, simulation should be used to supplement and check the adequacy of the finite-sample properties. It is important that the simulations mimic the actual data-generating processes.
- Statistical theory also has a role to guide the design of simulation studies and the analysis and presentation of simulation results. For example, it is immediately obvious that the scale parameter η need not be a factor in the experiment, as metrics of interest are invariant to the choice of η , even when data are censored. As mentioned in Section 7.1, under type 2 censoring and progressive failure censoring, RE comparing equivariant estimators of linear functions of the SEV location and scale parameters μ and σ will be invariant to both $\eta = \log(\eta)$ and $\beta = 1/\sigma$.
- When there are only a few failures, there is very little information in the data, as we have seen in our simulations. This lack of information is reflected by the extremely wide confidence intervals. In presenting results on an analysis, especially when data are limited (almost always the case) or there is potential for large losses if incorrect decisions are made, it is essential to quantify uncertainty as well as possible. The statistical uncertainty is relatively easy to quantify (with a confidence or prediction interval) and serves as a lower bound on the total uncertainty.

- When there are only a few failures, it may be necessary or desirable to supplement the data with external information. This is often done by assuming the value of the Weibull shape parameter, based on previous experience or knowledge of the physics of failure, and doing sensitivity analysis over a range of values. This approach is useful and is illustrated in Abernethy, Breneman, Medlin, and Reinman (1983), Nelson (1985), and Abernethy (2006). Some people refer to this approach as “Weibays”, but this is a confusing term because the method has no relationship to Bayesian methods and can be applied to the lognormal distribution just as readily as the Weibull distribution.

A useful alternative is to use a prior probability distribution to describe the uncertainty in β and do a Bayesian analysis, as illustrated in Chapter 14 of Meeker and Escobar (1998). This approach has the advantage of providing a point estimate and uncertainty interval, as in the classical approaches. The advantage of the sensitivity analysis approach is that it provides insight into which assumptions are conservative and which assumptions are not.

- In actual data-analysis and test planning applications, it is useful to use simulation to get insight into the properties of proposed tests and inference procedures. Plots of simulation results like those shown in Figure 9 allow an engineer or manager to clearly understand statements like “If the dark line is the truth, our estimates could be xx% off due to sampling variability.” It is clear that engineers and managers today have a much better understanding and appreciation for the role of variability. Both the popularity of Six-Sigma programs and the availability of powerful graphics/simulations tools have contributed to this.

There are several areas that need further research.

- Our study has focused on the Weibull distribution. It would be of interest to conduct similar studies for the other widely used distributions, especially the lognormal distribution. The results in Somboonsawatdee, Nair, and Sen (2007) suggest that there could be some interesting differences.
- Some previous work has been done on the reduction of bias in ML estimators (e.g., Thoman and Bain 1984 and Hiroshi 1999). Abernethy (2006) and Barringer (2009) also describe an approach to reduce the bias of ML estimators of the Weibull shape parameter. The effects of such efforts need to be evaluated using appropriate realistic censoring schemes and criteria for evaluating precision. Often efforts to reduce bias will result in increased MSE and this is not an improvement.
- The BLU and BLI estimators mentioned in Section 3.1 have optimality properties under type 2 censoring and could be expected to be approximately optimum for type 1 censoring, when $E(r)$ is large. It would be interesting to replicate our study, replacing MRR with BIE and BLU estimators to see how these procedure compare to ML estimators under type 1 censoring. These linear estimators will, however, also suffer, under type 1 censoring, because they also ignore information in the exact position of censored observations.

- The results from our simulation are conditional on having at least two failures. We did this to give MRR its best chance to performing well, as it has been suggested that MRR is better than ML in “small samples.” We have seen, however, that estimates (and especially MRR estimates) can take on extreme values when the number of failures is small (e.g., Figure 6). Thus it could be of interest to repeat our study, conditional on observing some larger number of failures (say three to ten). This would make sense if some alternative approach is to be used when the number of failures falls below a certain level.

Acknowledgments

We would like to thank Bob Abernethy for providing copies of *The New Weibull Handbook* to us and Paul Barringer for providing copies of Liu (1997) and results of his research to reduce bias in ML estimates. We also benefited from correspondence with Bob Abernethy, Luis Escobar, and Wes Fulton. We would also like to thank Luis Escobar, Katherine Meeker, Dan Nordman, and Yili Hong for providing helpful comments on an earlier version of this paper.

References

- Abernethy, R. B. (1996). *The New Weibull Handbook*, 2nd Edition. Robert B. Abernethy, 536 Oyster Road, North Palm Beach, FL 33408-4328.
- Abernethy, R. B. (2006). *The New Weibull Handbook*, 5th Edition. Robert B. Abernethy, 536 Oyster Road, North Palm Beach, FL 33408-4328.
- Abernethy, R. B., Breneman, J. E., Medlin, C. H., and Reinman, G. L. (1983). *Weibull Analysis Handbook*. Air Force Wright Aeronautical Laboratories Technical Report AFWAL-TR-83-2079. Available at <http://handle.dtic.mil/100.2/ADA143100>.
- Balakrishnan, N. and Aggarwala, R. (2000). *Progressive Censoring: Theory, Methods, and Applications* Boston: Birkhuser
- Barringer, P. (2009) Private communication.
- Bays, C. and Durham, S. (1976). Improving a Poor Random Number Generator, *ACM Transactions on Mathematical Software*, 2, 59-64.barringer
- Benard, A. and Bosi-Levenbach, E. C. (1953). The plotting of observations on probability paper, *Statistica Neerlandica*, 7, 163-173.
- Crowder, Kimber, Smith, Sweeting, (1991), *Statistical Analysis of Reliability Data*, New York: Chapman & Hall.
- Doganaksoy, N., Hahn, G.J., and Meeker, W.Q. (2000). Product Life Analysis: A Case Study, *Quality Progress* 33, 115-122.
- Escobar, L.A. (2009) Private communication.
- Genschel, U. and Meeker, W. Q. (2007). A comparison of maximum likelihood and median rank regression for weibull estimation, presented at the Joint Statistical Meetings, July 30, 2007.
- Gibbons, D. I. and Vance, L. C. (1981). A Simulation Study of Estimators for the 2-parameter Weibull Distribution. *IEEE Transactions on Reliability*, R-30, 61-66.

- Hahn, G.J., and Meeker, W.Q. (1991). *Statistical Intervals: A Guide for Practitioners*. John Wiley and Sons, Inc.
- Hampel, F. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* 69, 383-393
- Herd, G. R. (1960). Estimation of reliability from incomplete data, in *Proceedings of the 6th National Symposium on Reliability and Quality Control*, 202-217, New York: IEEE.
- Hiroshi, H. (1999). Bias Correction for Maximum Likelihood Estimates in the Two Parameter Weibull Distribution. *IEEE Transactions on Dielectrics and Electrical Insulation*. 6, 66-68.
- Hong, Y., Meeker, W.Q., and Escobar, L.A. (2008). The Relationship Between Confidence Intervals for Failure Probabilities and Life Time Quantiles, *IEEE Transactions on Reliability*, R-57, 260-266.
- Hong, Y., Meeker, W. Q., and McCalley, J. D. (2009). Prediction of Remaining Life of Power Transformers Based on Left Truncated and Right Censored Lifetime Data. *Annals of Applied Statistics* , 3 xxx-xxx (in press).
- Jeng, S. L. and Meeker W.Q. (2000). Comparisons of Weibull Distribution Approximate Confidence Intervals Procedures for Type I Censored Data. *Technometrics* 42, 135-148.
- Johnson, L. G. (1964). *The statistical treatment of fatigue experiments*, New York: Elsevier.
- Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, Second Edition, New York: John Wiley & Sons.
- Liu C.-C. (1997). *A Comparison Between the Weibull and Lognormal Models Used to Analyze Reliability Data*. Ph.D. Thesis, University of Nottingham. Available on line at <http://www.barringer1.com/wa.htm>.
- Mann, N. R., Schafer, R. E., and Singpurwalla, N. D. (1974). *Methods for Statistical Analysis of Reliability and Life Data*, New York: John Wiley & Sons.
- Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: John Wiley & Sons.
- Meeker, W. Q. and Escobar, L. A. (2004). *SPLIDA User's Manual*. Available from <http://www.public.iastate.edu/~splida/>.
- Nair, V. N. (1984). On the Behavior of Some Estimators from Probability Plots, *Journal of the American Statistical Association*, 79, 823-831.
- Nelson, W. (1969). Hazard Plotting for Incomplete Failure Data, *Journal of Quality Technology*, 1, 27-52.
- Nelson, W. (1982). *Applied Life Data Analysis*, New York: John Wiley & Sons.
- Nelson, W. (1985). Weibull Analysis of Reliability Data with Few or No Failures, *Journal of Quality Technology*, 17, 140-146.
- Nelson, W. (1990). *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses*, New York: John Wiley & Sons.
- Olteann, D. and Freeman, L. (2009). The Evaluation of Median Rank Regression and Maximum Likelihood Estimation Techniques for a Two-Parameter Weibull Distribution. *Quality Engineering*, xxx-xxx.
- Skinner, K. R., Keats, J. B., and Zimmer, W. J. (2001). A Comparison of Three Estimators

- of the Weibull Parameters, *Quality and Reliability Engineering International*, 17, 249-256.
- Somboonsavatdee, A., Nair, V. N., and Sen, A. (2007). Graphical Estimators from Probability Plots with Right-Censored Data, *Technometrics*, **49**, 420-429.
- Thoman, D. R. and Bain, L. J. (1984). Inferences on the Parameters of the Weibull Distribution, *Technometrics* 11, 445-460.
- Tobias, P. A., and Trindade, D. C. (1995). *Applied Reliability* (Second Edition), New York: Van Nostrand Reinhold.
- Vander Weil S., and Meeker, W.Q. (1990). Accuracy of Approximate Confidence Bounds Using Censored Weibull Regression Data from Accelerated Life Tests. *IEEE Transactions on Reliability* 39, 346-351.