9-20-2019

Coverage Properties of Weibull Prediction Interval Procedures to Contain a Future Number of Failures

Fanqi Meng
Iowa State University

William Q. Meeker
Iowa State University, wqmeeker@iastate.edu

Coverage Properties of Weibull Prediction Interval Procedures to Contain a Future Number of Failures

Abstract

Prediction intervals are needed to quantify prediction uncertainty in, for example, warranty prediction and prediction of other kinds of field failures. Naïve prediction intervals (also known as intervals from the "plug-in method") ignore the uncertainty in parameter estimates. Simulation-based calibration methods can be used to improve the accuracy of prediction interval coverage probabilities. This article investigates the finite-sample coverage probabilities for naive and calibrated prediction interval procedures for the number of future failures, based on the failure-time information obtained before a censoring time. We have designed and conducted a simulation experiment over combinations of factors with levels covering the ranges that are commonly encountered in practical applications. Our results indicate situations where the naïve prediction procedure performs poorly but where properly calibrated procedures do well. The simulation also uncovered exceptional cases, caused by the discreteness of the number of failures being predicted, where even the calibrated procedure can perform poorly.

Keywords

Censored data, coverage probability, maximum likelihood, prediction bound, simulation

Disciplines

Applied Statistics | Probability | Statistical Methodology | Statistical Theory

Coverage Properties of Weibull Prediction Interval Procedures to Contain a Future Number of Failures

Fanqi Meng and William Q. Meeker
Department of Statistics
Iowa State University
Ames, IA 50011
20 September 2019

Summary

Prediction intervals are needed to quantify prediction uncertainty in, for example, warranty prediction and prediction of other kinds of field failures. Naïve prediction intervals (also known as intervals from the "plug-in method") ignore the uncertainty in parameter estimates. Simulation-based calibration methods can be used to improve the accuracy of prediction interval coverage probabilities. This article investigates the finite-sample coverage probabilities for naive and calibrated prediction interval procedures for the number of future failures, based on the failure-time information obtained before a censoring time. We have designed and conducted a simulation experiment over combinations of factors with levels covering the ranges that are commonly encountered in practical applications. Our results indicate situations where the naïve prediction procedure performs poorly but where properly calibrated procedures do well. The simulation also uncovered exceptional cases, caused by the discreteness of the number of failures being predicted, where even the calibrated procedure can perform poorly.

Key Words

Censored data, coverage probability, maximum likelihood, prediction bound, simulation.

Acronyms

pdf	Probability density function
pmf	Probability mass function
cdf	Cumulative distribution function
BINCDF	Binomial cumulative distribution function
CI	Confidence interval
CP	Coverage probability
CCP	Conditional coverage probability
LPB	Lower prediction bound
PB	Prediction bound (could be lower or upper)

PI	Prediction interval (lower and upper together)
ML	Maximum likelihood
UCP	Unconditional coverage probability
UPB	Upper prediction bound

Notation

t_c Censoring time

t_w A future time point

E_r Expected number of failures before the censoring time t_c

E_m Expected number of failures between the censoring time t_c and t_w

K Number of failures between the censoring time t_c and t_w

$\underset{\sim}{K}$, $\tilde{K}$ Lower and upper prediction bounds, respectively, for K

$\underline{K}$, $\overline{K}$ Lower and upper calibrated prediction bounds, respectively, for K, when it is necessary to distinguish between naïve and calibrated intervals.

p_f Expected proportion failing before the censoring time t_c

n Sample size or number of units at the beginning of the test

1-α Nominal (desired) one-sided coverage probability

β, η Weibull shape and scale parameters, respectively

ρ Probability of failing between t_c and t_w, conditional on surviving until t_c

θ Parameter vector containing the Weibull parameters

^ Maximum likelihood estimate

I. Introduction

A. Objective

This work was motivated by several applications where it was required to predict the number of future field failures either for risk assessment (e.g., the number of failures that could result in serious consequences) or to determine the amount of reserves that were needed for future warranty claims. In all of these applications, because of the limited amount of data from the field and the need to report prediction intervals (PIs), special PI procedures, based on large-sample approximations, had to be employed. Although statistical theory tells us that the procedures used work well in large samples, there is a need to assess the adequacy of these approximations for finite samples over a range of practical situations. In this paper, we do this assessment by conducting a simulation study for the two most commonly used PI procedures.

Prediction of the number of failures in a given future time interval within a sample is a common practical problem. For example, as described in the section 12.5 of Meeker and Escobar (1998), a batch of product A had been inserted into the market, and a certain number of failed units had been returned for replacement before a data-freeze date (DFD). It was important for the finance department to have an accurate prediction for the number of additional failed units before the end of the warranty period, based on the existing product failure data. In other applications, safety was the primary concern. In addition to the point prediction, for many applications, it is also important to provide lower and upper bounds of the prediction (LPBs and UPBs), at some specified level of confidence.

A simple naïve method of obtaining PIs substitutes ML estimates for the true parameters. This naïve method uses appropriate quantiles of the distribution of the random variable to be predicted to define a PI and is asymptotically (as the expected number of failures approaches infinity) correct. While this naive procedure is generally simple to implement, it is possible for the actual CP to be poor, relative to the specified nominal CP (e.g., one might desire a 95% PI, but the actual probability that the PI will contain the future random quantity may be only 80%). It is, however, possible to use the concept of calibration to define a better PI procedure (i.e., one that has a coverage probability that is closer to the nominal confidence level).

In this study, we used a Monte Carlo simulation experiment to investigate the coverage properties of two PI procedures – the naïve method and the calibrated method. We evaluated CPs over the combinations of carefully-chosen experimental factors with levels covering the ranges commonly encountered in practical applications. Understanding these properties is important, so that analysts can choose the appropriate procedure to use and so that they understand the actual coverage probability that can be obtained.

B. Related Work

Considerable previous research has been done on the subject of statistical prediction. Simple exact methods are available for the normal distribution and complete data (e.g., Chapter 4 of Meeker, Hahn and Escobar 2017). For more complicated models and data, Cox (1975) suggested an analytical calibration method for obtaining the PIs, based on a Taylor series approximation and the asymptotic distribution of ML estimators. Today, calibration procedures are easier to implement by using simulation or bootstrap-based methods. Beran (1990) studied the large-sample asymptotic convergence properties of the CP of calibrated PI procedures. Lawless and Fredette (2005) introduced the concept of a predictive distribution, which provides an elegant approach for implementing the calibration procedure. Mee and Kushary (1994) showed how to use simulation to compute calibrated prediction intervals for Weibull failure times. Yang, See and Xie (2003) developed and evaluated a transformation to approximate normality approach to compute prediction intervals for Weibull failure times. Chapter 12 of Meeker and Escobar (1998), and Escobar and Meeker (1999) presented detailed methods of computing calibrated PIs for Weibull failure times and for the number of future Weibull failures, using life data, based on the large-sample approximate methods suggested in Cox (1975) and Beran (1990). Nordman and Meeker (2002) compared probability ratio, simplified probability ratio, and likelihood ratio Weibull prediction methods, assuming a known Weibull shape parameter values. The method of sub-sampling was presented by de Menezes et. al. (2006), which allowed one to compute a PI for a future number of failures with the knowledge only of the number of failures in the past time.

The finite-sample coverage properties of these prediction procedures have not been studied. Monte Carlo simulation provides a straight forward (albeit computationally intensive) method to evaluate the coverage probability properties of CI/PI procedures under realistic finite samples situations. Our focus is on prediction intervals for the number of future failures, as this is the most common application that we have encountered. For examples of such evaluations for other kinds of CIs and PIs, see Jeng and Meeker (2000), Nordman and Meeker (2002), and Genschel and Meeker (2010), and Chapters 6 and 7 of Meeker, Hahn, and Escobar (2017).

Hong and Meeker (2010) develop methods for predicting failures for individual failure modes for a product that fails in more than one way. Also, they show how use-rate information can be used to build a failure-time model based on use rather than time in service, eliminating extrapolation. Hong and Meeker (2013) extend the work of Hong and Meeker (2010) by allowing the use rate distribution of individual units to change over time. Xu et al. (2015) develop prediction methods for situations where there are delays in reporting and when there is a retirement distribution that, over time, depletes the number of units at risk to failure.

C. Overview

The rest of this paper is organized as follows. Section II describes the problem of predicting the number of additional future failures within a single sample, and the associated statistical model. Section III describes the methods of finding naïve and calibrated PIs for the within-sample prediction problem and introduces the concept of coverage probability (CP). Section IV presents an example to illustrate the computations and the use of naïve and calibrated PIs. Section V provides details of the design of the simulation experiment, and the procedure for the calculation of coverage probabilities for both PI methods. Section VI summarizes the general results from the simulation experiment. The final section gives concluding remarks and suggestions for some potential future research directions. The appendix explains some of the odd behavior in the CP results, caused by the discreteness effects of the Binomial distribution.

II. Prediction Model

A. Weibull Distribution

The well-known Weibull distribution cdf and pdf can be expressed, respectively, as

$$F_{Weibull}\left(t;\beta,\eta\right)=1-\exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right],$$

$$f_{Weibull}\left(t;\beta,\eta\right)=\frac{dF_{Weibull}\left(t;\beta,\eta\right)}{dt}=\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}\exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right],\ t>0$$

where $\beta>0$ is the shape parameter, and $\eta>0$ is the scale parameter (approximately the 0.632 quantile).

B. Within Sample Prediction Problem

Consider the situation in Figure 1 where n units start service at time point 0. The total number of failures by the censoring time t_c is r. The problem of interest is to use the information obtained up t_c to predict the number of additional failures K that will occur between t_c and a future time point t_w. For example, this t_w could be the end of a warranty period. Often, it is necessary to provide a PI to quantify the uncertainty in the point prediction.

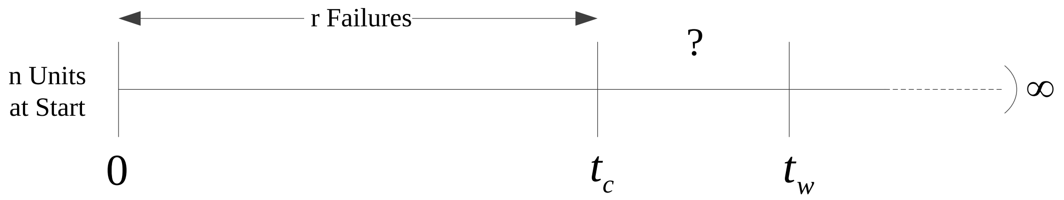


Figure 1. Within-Sample Prediction

Conditioning on r, the number of failures before t_c, K, the number of future failures between t_c and t_w, has a $Binomial(n-r, \rho)$ distribution, where

$$\rho = \Pr\left(\mathrm{t_c} < T \le t_w \mid T > \mathrm{t_c}\right) = \frac{\Pr\left(\mathrm{t_c} < T \le t_w\right)}{\Pr(T > \mathrm{t_c})} = \frac{F_{Weibull}\left(t_w;\beta,\eta\right) - F_{Weibull}\left(t_c;\beta,\eta\right)}{1 - F_{Weibull}\left(t_c;\beta,\eta\right)}. \quad (1)$$

This quantity is the probability that a given unit will fail in the interval $\left(t_c,\ t_w\right)$, given that the unit has survived to time t_c.

III. Statistical Prediction Interval Procedures and Their Coverage Probability

A. Naïve Prediction Intervals

The purpose of a within-sample statistical PI is to predict the additional failures K, in the future time interval of $\left(t_c,\ t_w\right)$, and produce a PI denoted by $[\underset{\sim}{K},\ \tilde{K}]$, based on the failure-time data obtained before the time point t_c. A simple one-sided approximate $100\left(1-\alpha\right)\%$ naïve LPB and a corresponding $100\left(1-\alpha\right)\%$ UPB for K are obtained, respectively, as the α and $1-\alpha$ quantiles of the $Binomial\left(n-r, \hat{\rho}\right)$ distribution, where the ML estimate of ρ, $\hat{\rho}$ is calculated according to (1), by directly substituting the ML estimates of β and η, obtained from the failure-time data (i.e, the failures before the censoring time t_c). This is also known as the “plug-in” method of computing LPBs and UPBs. Together these bounds provide an approximate $100\left(1-2\alpha\right)\%$ two-sided prediction interval (PI). The actual coverage probability of this naïve procedure cannot be expected to be close to the nominal $100\left(1-2\alpha\right)\%$.

B. Coverage Probabilities for a Statistical Prediction Interval Procedure

Here we review some of the basic ideas given in Escobar and Meeker (1999), as these are needed in our study. Generally, a statistical PI procedure is designed to have a given nominal CP (i.e., probability of capturing the random variable to be predicted). In many applications, it is impossible to find a prediction procedure that will have exactly the desired CP and the actual CP will be only approximately equal to the nominal CP.

Technically, there are two types of CPs, and it is important to distinguish between them.

1. Conditioning on the number of failures r in $\left(0,\ t_c\right)$ and the estimates of Weibull parameters, the number of additional failures K has a Binomial distribution,

and the resulting PI with nominal confidence level $1-2\alpha$ has the conditional coverage probability (CCP):

$$\begin{aligned} CCP[PI(1-2\alpha)\mid r;\hat{\beta},\hat{\eta}] &= \Pr(\underset{\sim}{K} \le K \le \tilde{K}) \\ &= BINCDF(\tilde{K}, n-r, \hat{\rho}) - BINCDF(\underset{\sim}{K}-1, n-r, \hat{\rho}) \end{aligned} \quad (2)$$

Note that the CCP of a particular PI is random because the PI $[\underset{\sim}{K},\ \tilde{K}]$ depends on the data through the Weibull parameter estimates and the number of survivors r , which vary from sample to sample. For any particular PI, the CCP is also unknown because it depends on the actual but unknown Weibull parameters. For these reasons, the CPP cannot be used, directly, to evaluate the properties of a PI procedure.

2. The unconditional coverage probability (UCP) is used to evaluate a PI procedure. The UCP is obtained as the expectation of the CCP over all possible samples,

$$\begin{aligned} UCP[PI(1-2\alpha);\beta,\eta] &= \Pr(\underset{\sim}{K} \le K \le \tilde{K};\beta,\eta) \\ &= E\left\{CCP[PI(1-2\alpha)\mid r;\hat{\beta},\hat{\eta}]\right\} \end{aligned} \quad (3)$$

When the meaning is clear from the context, we will, as is commonly done, use CP to refer to UCP.

C. Calibrated Prediction Intervals

The basic idea of calibration can be illustrated with the following example. Suppose a 95% PI is desired, but simulation shows that the naïve procedure has a CP of only 0.90. One might expect that changing the input request to something like 97% or 98% would result to an actual CP closer to the nominal confidence level of 95%. The LPB and the UPB are calibrated separately. As suggested by Cox (1975), the naïve LPBs and UPBs can be calibrated by finding, respectively, values $1-\alpha_{cl}$ and $1-\alpha_{cu}$ such that both the α_{cl} and the $1-\alpha_{cu}$ $Binomial\left(n-r,\hat{\rho}\right)$ quantiles have an UCP that is at least $1-\alpha$. That is, the calibration procedure determines the input confidence levels $1-\alpha_{cl}$ and $1-\alpha_{cu}$ such that the simulation evaluation CP will be approximately equal to the desired level. Calibrated PI procedures for predicting discrete random variable, based on censored data are still approximate because the random variable being predicted are discrete, and because the actual UCP of the procedure depends on the unknown lifetime distribution parameters.

Cox's analytical calibration method is based on a Taylor series approximation for the asymptotic properties of the maximum likelihood estimators. Generally, conducting analytical calibration is only possible for some simple cases. Today, calibration is more easily done by simulation of the sampling-prediction process. Steps 4.1 – 4.8 in the overall algorithm presented in Section V, describes the detailed calibration procedure.

D. Relationship between One-sided and Two-sided Prediction Intervals

Combining a one-sided $100(1-\alpha_1)\%$ LPB and a one-sided $100(1-\alpha_2)\%$ UPB can be used to generate a two-sided $100(1-\alpha_1-\alpha_2)\%$ PI. Usually, analysts want to know how good things might be as well as how bad thing might be. Although it is theoretically possible to determine the values of α_1 and α_2 to minimize the expected length of a two-sided PI, analysts tend to expect and find it easier to interpret an equal tail ($\alpha_1=\alpha_2$) two-sided PI. This is because the endpoints of the PI can more easily be interpreted as one-sided PBs. Being able to interpret the endpoints of a PI as two separate one-sided PBs is important because, for example, the cost of underprediction of the future number of failures may be far more than the cost of over prediction. Thus, in this paper, we focus on the coverage properties of one-sided PBs.

IV. Numerical Example: Prediction Interval to Contain the Number of Future Product-A Failures

This example, originally presented in Meeker and Escobar (1998), illustrates the application and computation of the PI procedures being evaluated here. During one month, $n=10,000$ units of Product-A (the actual name of the product is not being used to protect sensitive information) were put into service. After 48 months, 80 failures had been reported. Management requested a point prediction and a UPB on the number of the remaining $n-r=10,000-80=9920$ units that will fail during the next 12 months (i.e., between 48 and 60 months of age). We will also provide an LPB. The available data and previous experience suggested a Weibull failure-time distribution, and the ML estimates are $\hat{\eta}=1{,}152$ and $\hat{\beta}=1.518$. From these,

$$\hat{\rho}=\frac{\hat{F}_{Weibull}(60)-\hat{F}_{Weibull}(48)}{1-\hat{F}_{Weibull}(48)}=0.003233$$

The point prediction for the number failing between 48 and 60 months is $\hat{K}=(n-r)\times\hat{\rho}=32.07$. The naïve 95% UPB on K is $\tilde{K}(0.95)=K_{0.95}=42$, the smallest integer k such that $BINCDF(k,9920,0.003233)\geq 0.95$. The calibration curve shown in Figure 2 was obtained by using the simulation procedure given in Escobar and Meeker (1999) (and also described in Section IV of this paper). For a nominal 0.95 CP this curve suggests that using $1-\alpha_{cu}=0.986$ as the input confidence level for the UPB should yield a $UCP[PI(0.986);\hat{\boldsymbol{\theta}}]\cong 0.95$, Thus, the calibrated approximate 95% UPB on K is $\tilde{K}(0.986)=K_{0.986}=45$, the smallest integer k such that $BINCDF(k,9920,0.003233)\geq 0.986$.

The naïve 95% LPB on K is $\underset{\sim}{K}(0.95)=K_{0.05}=22$, the largest integer k such that $BINCDF(k,9920,0.003233)\leq 0.05$. Similarly, the calibration curve shown in Figure 2 gives, $1-\alpha_{cl}=0.981$ for the input confidence level for the LPB, that should yield $UCP[PI(0.981);\hat{\boldsymbol{\theta}}] \cong 0.95$. Thus, the calibrated approximate 95% LPB on K is $\underset{\sim}{K}(0.981) = K_{0.019} = 20$, the largest integer k such that $BINCDF(k, 9920, 0.003233) \leq 0.019$.

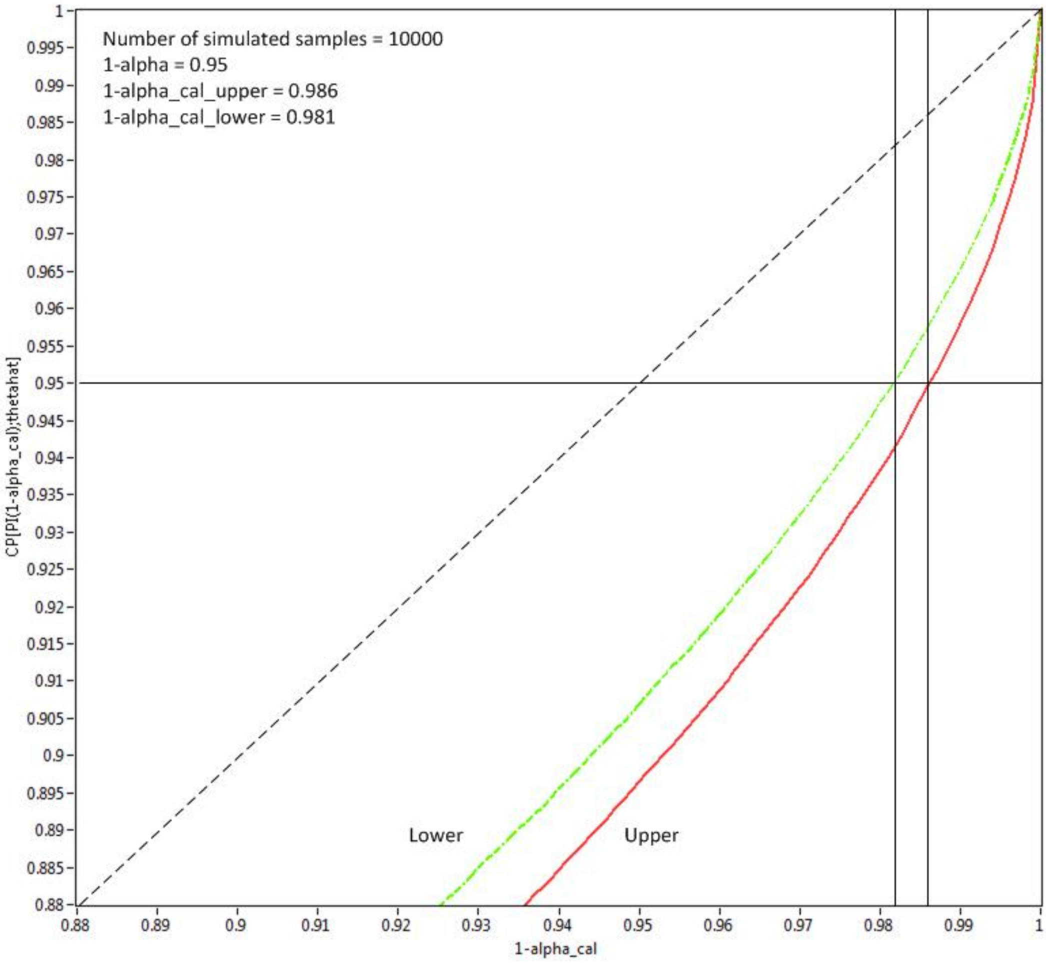


Figure 2. Calibration Curves for Upper and Lower Prediction Bounds on the Number of Field Functions in the Next Year for the Product-A Population.

Although the large sample approximation theory suggests that the PI obtained by the calibrated method should have a better CP, in this finite sample the actual coverage probabilities for both procedures still remain unknown. Thus, it is necessary to assess them by using a simulation experiment.

V. Simulation Experiment

A. Experiment Design

The following experimental factors were selected to study the CP of both the naïve and the calibrated PI procedures.

p_f : the expected proportion failing before the censoring time t_c

E_r : the expected number of failures before the censoring time t_c

E_m : the expected number of failures between the censoring time t_c and a future time t_w

β : the Weibull shape parameter

Without loss of generality, η is always taken to be 1, because it is a scale parameter for Weibull distribution.

E_r was chosen as a factor because the amount of information in a censored sample is roughly proportional to the number of failures. Also, as was seen in Jeng and Meeker (2000), the combination of factors E_r and p_f will have little interaction effect (as opposed to using p_f and the sample size n), making interpretation of the simulation results much simpler. The first three factors have to be converted to another set of factors to correspond to the data to be simulated from our statistical model. The converted factors are:

n : the sample size or number of units at the beginning of the test.

t_c : the censoring time

t_w : the future time point

The formulas for the conversions are:

$$n = E_r / p_f$$

$$t_c = F^{-1}_{Weibull}(p_f; \beta, \eta)$$

$$t_w = F^{-1}_{Weibull}(p_m; \beta, \eta), \text{ where } p_m = (E_r + E_m)/n$$

These conversions are subject to the following constraints on the original set of factors:

n should be an integer;

$E_r + E_m \le E_r / p_f$, which can also be expressed as $E_m \le E_r \left(1/p_f - 1\right)$

The levels of the factors used in our simulation experiment were:

p_f : 0.01, 0.05, 0.1, 0.3, and 0.5

β : 0.8, 1.0, 1.5, and 3.0

E_r : 5, 10, and 20

E_m : 0.5, 1, 3, 5, 7, 10, 20, and 50 (where the level is valid only if it is less than the upper limit given in Table 1 for combinations of E_r and p_f , because E_m must be less than the expected size of the risk set.)

Table 1. The upper limit for E_m for all the combinations of p_f and E_r

p_f

		0.01	0.05	0.1	0.3	0.5
	5	495	95	45	11	5
E_r	10	990	190	90	23	10
	20	1980	380	180	46	20

B. Evaluation of Prediction Interval Procedures

Corresponding to actual applications we have encountered, our simulation is based on Type I (or time censoring). With Type I censoring there is always a positive probability that there will be 0 failures and in such cases, the ML estimates of the Weibull parameters do not exist. ML estimates based on only one failure are not very useful because they are subject to a large amount of statistical uncertainty. As done in other Type I censoring simulations (e.g., Genschel and Meeker 2010, and Jeng and Meeker 2000), we use only valid samples defined as simulated samples with at least two failures. Thus our results are conditional on the event of having two or more failures. Table 2 gives the probability of having an invalid sample for each of the combinations of the experimental factor levels. We can see that these probabilities are negligible unless E_r is small.

		p_f				
		0.01	0.05	0.1	0.3	0.5
	5	0.04	0.04	0.03	0.02	0.01
E_r	10	0.0005	0.0004	0.0003	8×10^{-5}	2×10^{-5}
	20	4×10^{-8}	3×10^{-8}	2×10^{-8}	1×10^{-9}	4×10^{-11}

Table 2 Probabilities of invalid samples

The precise steps of our simulation algorithm are

1. Using the true parameters $\boldsymbol{\theta}$, simulate B_1 valid samples according to the pre-specified censoring pattern.
2. For each sample i, $i = 1,...,B_1$, compute $\hat{\boldsymbol{\theta}}_i$, the ML estimate of $\boldsymbol{\theta}$ based on sample i, and the number of failures r_i. Then compute the Binomial parameters $(n - r_i)$, and

$$\hat{\rho}_i = \frac{F(t_w;\hat{\boldsymbol{\theta}}_i) - F(t_c;\hat{\boldsymbol{\theta}}_i)}{1 - F(t_c;\hat{\boldsymbol{\theta}}_i)} \quad . \tag{4}$$

3. For each sample i, $i = 1,..., B_1$, compute the naïve $1-\alpha$ UPB and LPBs $\tilde{K}_i$ and $\underset{\sim}{K}_i$, where $\tilde{K}_i$ is the smallest K such that $BINCDF(K; n - r_i, \hat{\rho}_i) \geq 1-\alpha$, and $\underset{\sim}{K}_i$ is the largest K such that $BINCDF(K; n - r_i, \hat{\rho}_i) < \alpha$;
4. For each sample i, $i = 1,..., B_1$, compute the calibrated $1-\alpha$ UPBs and LPBs $\overline{K}_i$ and $\underline{K}_i$, where $\overline{K}_i$ is the smallest k such that $BINCDF(k; n - r_i, \hat{\rho}_i) \geq 1-\alpha_{cu}$, and $\underline{K}_i$ is the largest k such that $BINCDF(k; n - r_i, \hat{\rho}_i) < 1-\alpha_{cl}$, where α_{cu} and α_{cl} are obtained according to the following procedure:
   4.1. Using $\hat{\boldsymbol{\theta}}_i$, simulate B_2 valid samples according to the pre-specified censoring pattern;
   4.2. For each sample j in $j = 1,..., B_2$, obtain the ML estimate ($\hat{\boldsymbol{\theta}}_j$), the number of failures r_j, and Binomial parameters $n - r_j$, and
   $$\hat{\rho}_j = \frac{F(t_w; \hat{\boldsymbol{\theta}}_j) - F(t_c; \hat{\boldsymbol{\theta}}_j)}{1 - F(t_c; \hat{\boldsymbol{\theta}}_j)} \quad . \tag{5}$$
   4.3. Choose a value of α, say, α_0;
   4.4. For each sample j in $j = 1,..., B_2$, obtain the naïve $1-\alpha_0$ UPB and LPB $\tilde{K}_j$ and $\underset{\sim}{K}_j$, where $\tilde{K}_j$ is the smallest k such that $BINCDF(k; n - r_j, \hat{\rho}_j) \geq 1-\alpha_0$, and $\underset{\sim}{K}_j$ is the largest k such that $BINCDF(k; n - r_j, \hat{\rho}_j) < \alpha_0$;
   4.5. For each sample j in $j = 1,..., B_2$, evaluate CCPs $P_{nuj} = BINCDF(\tilde{K}_j; n - r_j, \hat{\rho}_i)$, and $P_{nlj} = 1 - BINCDF(\underset{\sim}{K}_j - 1; n - r_j, \hat{\rho}_i)$;
   4.6. The UCP of calibrated UPB and LPB are then, respectively, $\left(\sum_{j=1}^{B_2} P_{nuj}\right) \Big/ B_2$ and $\left(\sum_{j=1}^{B_2} P_{nlj}\right) \Big/ B_2$ for the nominal level of $1-\alpha_0$;
   4.7. Repeat steps 4.3 to 4.6, using a large number of consecutive values of $1-\alpha_0$ between 0 and 1, and obtain the corresponding UCPs for the calibrated LPBs and UPBs, giving the calibration curve.
   4.8. Use linear interpolation to find the desired UCPs, find the corresponding values, say, $1-\alpha_0 = 1-\alpha_{cu}$ and $1-\alpha_0 = 1-\alpha_{cl}$ on the calibration curves.
5. For each sample $i = 1,..., B_1$, evaluate the naïve PI CCPs $P_{nui} = BINCDF(\tilde{K}_i; n - r_i, \rho)$, and $P_{nli} = 1 - BINCDF(\underset{\sim}{K}_i - 1; n - r_i, \rho)$, where
   $$\rho = \frac{F(t_w; \boldsymbol{\theta}) - F(t_c; \boldsymbol{\theta})}{1 - F(t_c; \boldsymbol{\theta})} \tag{6}$$
6. The UCP of naïve UPB and LPB are $\left(\sum_{i=1}^{B_1} P_{nui}\right) \Big/ B_1$, and $\left(\sum_{i=1}^{B_1} P_{nli}\right) \Big/ B_1$;

7. For each sample i, $i = 1,...,B_1$, evaluate the calibrated PI CCPs $P_{cui} = BINCDF(\overline{K}_i; n - r_i, \rho)$, and $P_{cli} = 1 - BINCDF(\underline{K}_i - 1; n - r_i, \rho)$;
8. The UCP of the calibrated UPB and LPB procedures are then $\left(\sum_{i=1}^{B_1} P_{cui}\right) \Big/ B_1$, and $\left(\sum_{i=1}^{B_1} P_{cli}\right) \Big/ B_1$, respectively.

In this simulation experiment, B_1 and B_2 were both chosen to be 2000. These numbers were chosen as a compromise to control the amount of computer time required for all of the runs and to keep the Monte Carlo error in the simulation results small. We conducted a pilot study to investigate the Monte Carlo standard errors of the UCPs and observed that the resulting standard errors of the coverage probabilities varied as a function of the factor level combinations but were always smaller than 0.01 for the chosen values of B_1 and B_2.

VI. Simulation Experiment Results

In this section, we present the most interesting and useful results of the simulation experiment. Here, only the results for $\beta = 1$ are reported, because no difference in any results could be detected across the different levels of β values. A similar approximate invariance property was also reported in Genschel and Meeker (2010), where the simulation study to answer a different question was also conducted under a Type I censoring scheme. For Type II censoring, Escobar (2010) provided an analytical demonstration that the distribution of properly scaled ML estimates are invariant of the choice of β value, and this explains why it is that β has no discernable effect in the results of our simulation.

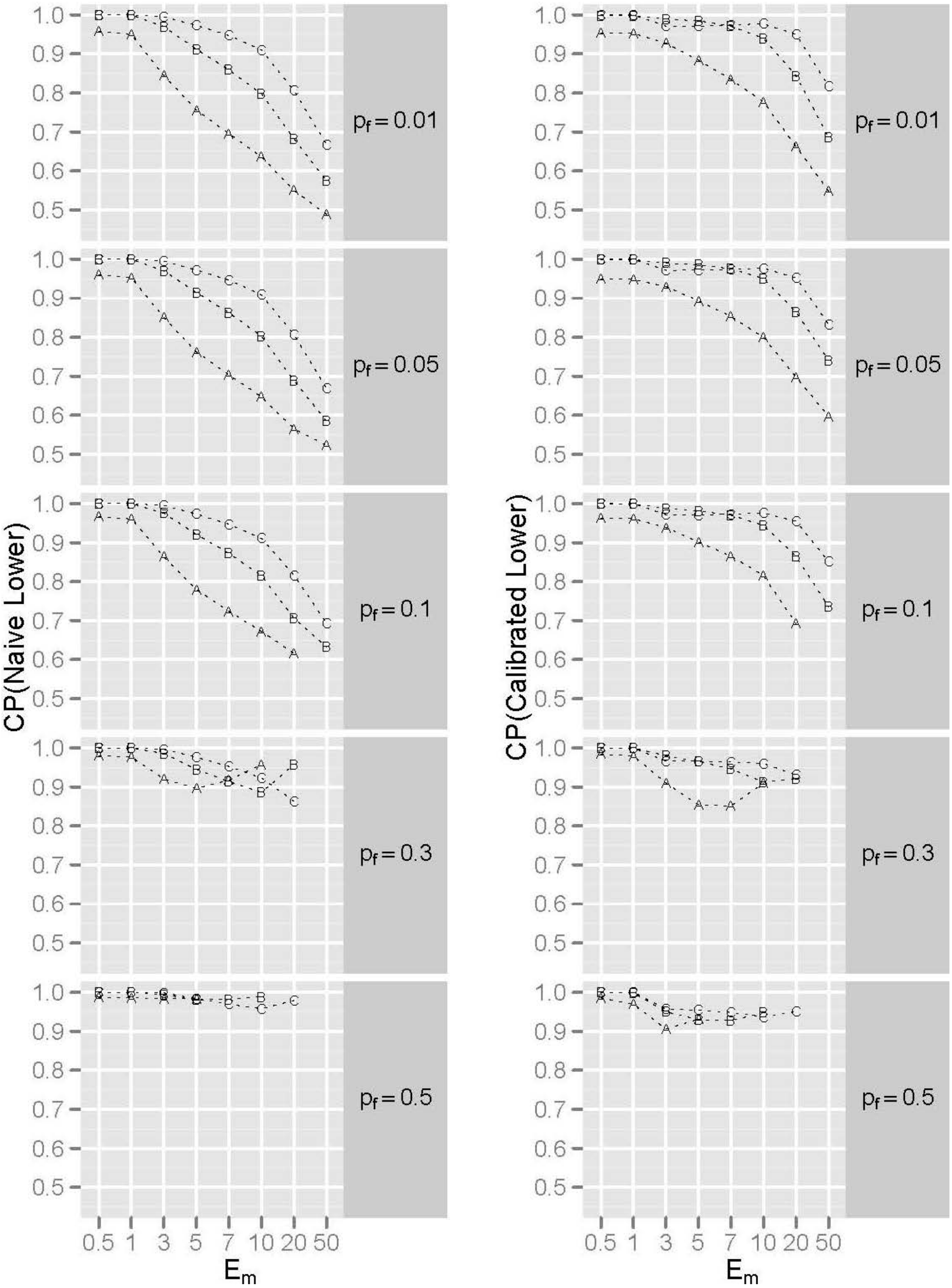


Figure 3. CPs versus E_m [the expected number of future failures in the time interval (t_c, t_w)] for the naïve and calibrated LPBs. The letters (A, B, C) in each plot correspond to $E_r = (5, 10, 20)$ [the expected number failing in the time interval $(0, t_c)$].

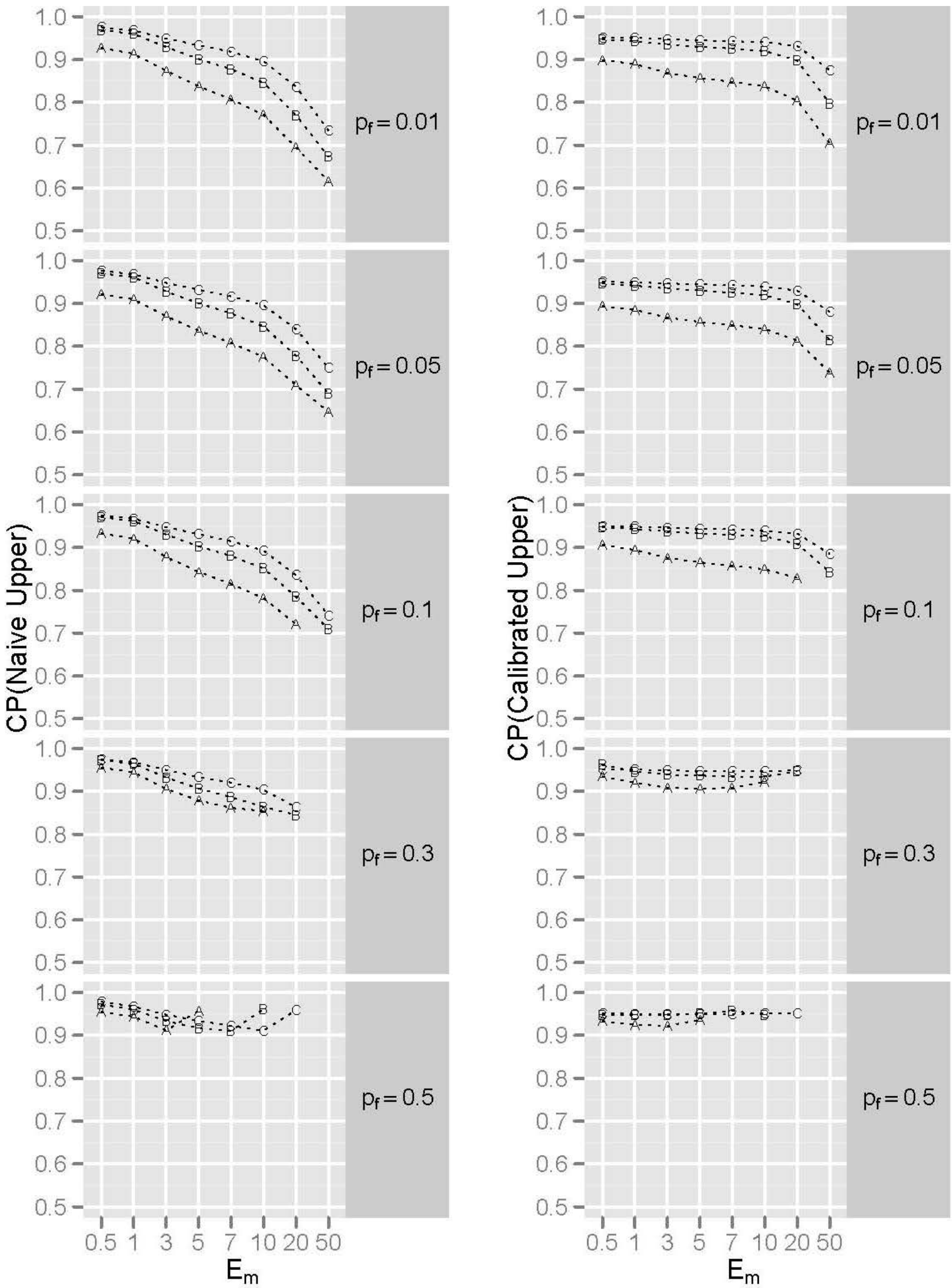


Figure 4. CPs versus E_m [the expected number of future failures in the time interval (t_c, t_w)] for the naïve and calibrated UPBs. The letters (A, B, C) in the lines of each plot correspond to $E_r = (5, 10, 20)$ [the expected number failing in the time interval $(0, t_c)$].

Figure 3 shows the CP of the one-sided 95% naive and calibrated LPBs. It can be seen from this figure that for small values of p_f (0.01~0.1), the CPs deviate below the nominal 95% level as E_m increases. This observation reveals that larger extrapolation in prediction time will cause the CP to be seriously below the nominal level. Also, larger values of E_r provide better CPs at the same level of E_m. This is because large E_r values typically result in more failures before the censoring time, which provides more accurate estimation of the distribution parameters and better accuracy for the large-sample approximations.

The calibration method usually (but not always) provides a CP that is closer to nominal, when compared with the naïve method. As seen in Figure 3, for small values of p_f, the CPs for the calibrated method are consistently greater than the ones for the naive method. However, for small to moderate values of E_m, the calibration procedure tends to overcorrect the CPs, making them close to 1. This is due to the discreteness of the binomial distribution and the conservative nature of the rule used to compute the naïve PBs. Details of this behavior are described in the Appendix.

For large values of p_f (0.3, 0.5) and small E_r, the CPs tend to oscillate as a function of E_m. This fluctuation is again due to the discreteness of the binomial distribution of the future number of failures, conditional on the number of failures before the censoring time. Agresti and Coull (1998) and Chapters 6 and 7 of Meeker, Hahn and Escobar (2017) reported similar patterns in the plots of CPs for other kinds of statistical intervals (CIs and PIs, respectively) where the Binomial distribution was involved.

Plots of CPs for the UPBs in Figure 4 show that the calibrated method provides consistently better coverage properties for $E_r \geq 10$. It is seen that although the CPs decrease as the value of E_m increases, the speed of the decrease is slower for the calibrated UPBs. When $E_m \leq 20$, the CPs are generally above 90% for the $E_r = 10, 20$ cases. In addition to the effect of increasing CPs that were below the nominal level, the calibration method also reduces the conservative coverage probabilities so that they are closer to the nominal level. This effect is seen for $E_r = 10, 20$ and $E_m = 0.5$ cases, where the CPs for the UPBs constructed by the naive method were close to 1, while the calibrated method brought the CPs close to the 95% level.

VII. Conclusion, Summary, and Directions for Future Research

We have investigated the properties of two important PI procedures for predicting a future number of failures of a product in the field. Because of the Type I censored data and a discrete random variable being predicted in these applications, there is no known exact PI procedure. Thus large sample approximate prediction methods need to be used. It is important to know when the large-sample approximate methods provide adequate approximations. We compared the naive method and the calibrated method of constructing the needed PIs. We utilized Monte Carlo simulations to evaluate the CPs and compare them with the commonly-used nominal 0.95 CP. Our results can be summarized as follows:

1. The calibrated PI procedure usually, but not always provides a closer-to-nominal level CP, when compared with the naive PI procedure.

2. A moderately large number of failures (say 10 to 20) before the censoring time are needed, in order to guarantee an adequate CP, even for the calibrated procedure.

3. Because of the discreteness of the Binomial distribution and the usual definition of binomial quantiles used in the prediction procedures, the CPs are generally conservative when E_m is close to 0 or close to the size of the risk set (the two extreme possibilities for the binomial outcome).

4. When E_m is not extreme, the CP is a decreasing function of E_m. This implies that the performance of the procedure degrades as more and more extrapolation is used to make the prediction. The PI obtained from the calibrated method decreases more slowly when compared with the PI obtained from the naive method.

Although our study has focused on PIs for additional failures from a single group, the approach could be applied to more complicated PI evaluation problems, for example, the problem of prediction of future failures from multiple groups of units with staggered entry into field (as described in Escobar and Meeker 1999). Prediction procedures for other kinds of applications have been provided in previous work such as Escobar and Meeker (1999). It would be useful to evaluate these PI procedures, not only with the purpose of verifying the coverage properties, but also to provide some guidance (for example, determination of the amount of data that is required for adequate approximations to be available) on the planning of reliability studies, based on the distribution of the prediction width.

Appendix

This appendix describes the special properties of the Binomial distribution, and shows analytically how the discreteness affects the CPs of PI procedures for the future number of failures. First, we note that when ρ is known or if the sample size is large enough that $\hat{\rho}$ is close to ρ, (11) will provide a conservative PI, due to the usual definition of binomial quantiles used in these prediction procedures (the quantile of a discrete random variable with cdf $F(x)$ is defined as the smallest x such that $F(x) \geq p$). In the rest of this appendix, we first consider situations when $\hat{\rho}$ is close to zero (corresponding to a small predicted number of failures). Then we consider the situation where $\hat{\rho}$ is close to 1 (corresponding to situations when almost all of the units in the risk set will fail).

As can be seen from Figure 3, the CPs for the naive procedure are almost always equal to 1 when $E_r = 10, 20$ and $E_m = 0.5$. This special effect can be explained by the following analysis.

The UCP for the naïve LB is given by

$$\begin{aligned} UCP[PI(1-\alpha_0);\boldsymbol{\theta}] &= E_r\{CP[PI(1-\alpha_0) \mid r;\boldsymbol{\theta}]\} \\ &= E_r[1-BINCDF(\underset{\sim}{K}_r - 1); n-r, \rho] \end{aligned} \tag{7}$$

where

$$\begin{aligned} \rho &= \frac{F_{Weibull}(t_w,\boldsymbol{\theta}) - F_{Weibull}(t_c,\boldsymbol{\theta})}{1-F_{Weibull}(t_c,\boldsymbol{\theta})} = \frac{\dfrac{E_r+E_m}{n} - \dfrac{E_r}{n}}{1-\dfrac{E_r}{n}} \\ &= \frac{E_m}{n-E_r} = \frac{E_m}{\dfrac{E_r}{p_f} - E_r} = \frac{E_m}{E_r} \cdot \frac{p_f}{1-p_f} \end{aligned} \tag{8}$$

The Monte Carlo evaluation of (8) is given by

$$UCP[PI(1-\alpha_0);\boldsymbol{\theta}] \approx \frac{\sum_{i=1}^{B_1} P_i}{B_1} \tag{9}$$

where

$$P_i = CP[PI(1-\alpha_0) \mid r_i;\boldsymbol{\theta}] = 1 - BINCDF(\underset{\sim}{k}_i - 1; n-r_i, \rho) \tag{10}$$

where $\underset{\sim}{k}_i$ is selected as the largest integer K, such that

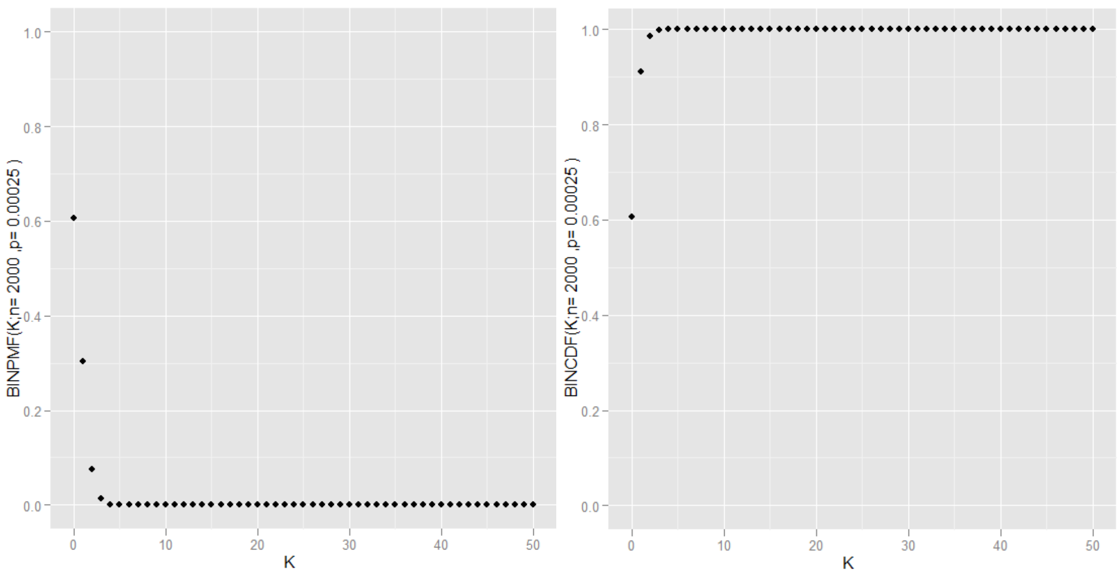
$$BINCDF(K; n-r_i, \hat{\rho}_i) < \alpha_0 \tag{11}$$

within the simulated sample i. With a moderate number of failures (i.e. $E_r = 10$ and $E_r = 20$ cases), it is reasonable to assume that the ML estimate of ρ will be close to the true value. Therefore, we have

$$\hat{\rho}_i \approx \frac{E_m}{E_r} \cdot \frac{p_f}{1-p_f} \tag{12}$$

Because $f(p_f) = p_f \big/ (1-p_f)$ is a monotonically increasing function of p_f, from (13), it can be inferred that with a small ratio of E_m / E_r and a small value of p_f, the resulting $\hat{\rho}_i$ will be small.

As an example, for the extreme case $E_m = 0.5$, $E_r = 20$, and $p_f = 0.01$, the mass of the sampling distribution $\hat{\rho}_i$ will concentrate around 0.00025. Figure 5 for K shows that with such a small value of $\hat{\rho}_i$, the probability mass of the binomial distribution is primarily located in the region where K is small. e.g., $\Pr(K = 0) \cong 0.6$. Thus when selecting UPB $\underset{\sim}{K}_i$ values using (11), such a combination of parameters (a small ratio of E_m / E_r and a small value of p_f) will almost always provide a LPB $\underset{\sim}{K}_i = 0$. This LPB will have a CCP equal to 1, when evaluated by (7), regardless of the true parameter value of ρ.



(a) (b)

Figure 5. Binomial pmf(a) and cdf(b) for $n = 2000$ and $\rho = 0.00025$

There is a similar explanation for why the CPs approach 1 when E_m is large in the case of $p_f = 0.5$, as seen in Figure 4. For this case, $E_m \approx n - E_r = (E_r / p_f) - E_r$. Because the large ratio of E_m / E_r and large value of p_f results in a large $\hat{\rho}_i$ ($\hat{\rho}_i \cong 1$), most of the mass in the Binomial pmf is close to $n-r$ (the largest possible number of future failures). In such cases, the UPB will almost always be equal to $\tilde{K}_i = n-r$, causing the LPBs to be extremely conservative. Because the value for the obtained bound is not always at the extreme, however, its CCP also depends on the true parameter value ρ. Thus, the variability in estimating $\hat{\rho}_i$ will make the UCP slightly smaller than one.

References

A. Agresti and B. Coull (1998), "Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119-126.

R. Beran (1990), "Calibrating Prediction Regions," *Journal of the American Statistical Association*, 85, 715-723.

D. R. Cox (1975), "Prediction Intervals and Empirical Bayes Confidence Intervals," *Perspectives in Probability and Statistics*, ed. J. Gani, London: Academic Press, 47-55.

F. S. de Menezes, M. J. Ferrua Vivanco and L. C. Sampaio (2006), "Determination of Prediction Intervals for a Future Number of Failures: A Statistical and Monte Carlo Approach," *Brazilian Journal of Physics*, 690-699.

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

L. A. Escobar (2010), "Comments on Genschel and Meeker," *Quality Engineering*, 22, 284-288.

U. Genschel and W. Q. Meeker (2010), "A Comparison of Maximum Likelihood and Median-Rank Regression for Weibull Estimation," *Quality Engineering*, 22, 236-255.

Y. Hong, and W. Q. Meeker (2010). Field-failure and warranty prediction based on auxiliary use-rate information. *Technometrics* 52, 148–159.

Y. Hong, and W. Q. Meeker (2013). Field-failure predictions based on failure-time data with dynamic covariate information. *Technometrics* 55, 135–149.

Y. Hong, 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.

S. L. Jeng and W. Q. Meeker (2000), "Comparisons of Weibull Distribution Approximate Confidence Intervals Procedures for Type I Censored Data," *Technometrics*, 42, 135-148.

J. F. Lawless and M. Fredette (2005), "Frequentist Prediction Intervals and Predictive Distributions," *Biometrika*, 92, 3, 529-542.

R. Mee and D. Kushary (1994), "Prediction Limits for the Weibull Distribution Utilizing Simulation," *Computational Statistics and Data Analysis*, 17, 327-336.

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

W. Q. Meeker, G.J. Hahn, and L.A. Escobar, (2017), Statistical Intervals: A Guide for Practitioners and Researchers, Second Edition. John Wiley and Sons, Inc.

D. Nordman and W. Q. Meeker (2002), "Weibull Prediction Intervals for a Future Number of Failures," *Technometrics*, 44, 15-23.

Z. Xu, Y. Hong, and W. Q. Meeker (2015). Assessing risk of a serious failure mode based on limited field data. *IEEE Transactions on Reliability* 64, 51–62.

Z. Yang, S. P. See and M. Xie (2003), "Transformation Approaches for the Construction of Weibull Prediction Interval," *Computational Statistics and Data Analysis*, 43, 357-368.