10-2014

Simultaneous Prediction Intervals for the (Log)-Location-Scale Family of Distributions

Yimeng Xie
Virginia Tech

Yili Hong
Virginia Tech

Luis A. Escobar
Louisiana State University

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

Simultaneous Prediction Intervals for the (Log)-Location-Scale Family of Distributions

Abstract
Making predictions of future realized values of random variables based on currently available data is a frequent task in statistical applications. In some applications, the interest is to obtain a two-sided simultaneous prediction interval (SPI) to contain at least k out of m future observations with a certain confidence level based on n previous observations from the same distribution. A closely related problem is to obtain a onesided upper (or lower) simultaneous prediction bound (SPB) to exceed (or be exceeded) by at least k out of m future observations. In this paper, we provide a general approach for constructing SPIs and SPBs based on data from a member of the (log)-location-scale family of distributions with complete or right censored data. The proposed simulationbased procedure can provide exact coverage probability for complete and Type II censored data. For Type I censored data, the simulation results show that our procedure provides satisfactory results in small samples. We use three applications to illustrate the proposed simultaneous prediction intervals and bounds.

Keywords
censored data, coverage probability, k out of m, lognormal, simulation, Weibull

Disciplines
Statistics and Probability

Comments
This preprint was published as Yimeng Xie, Yili Hong, Luis A. Escobar, and William Q. Meeker, "Simultaneous Prediction Intervals for the (Log)-Location-Scale Family of Distributions".

Simultaneous Prediction Intervals for the (Log)-Location-Scale Family of Distributions

Yimeng Xie and Yili Hong
Department of Statistics
Virginia Tech
Blacksburg, VA 24061

Luis A. Escobar
Department of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70803

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

Abstract

Making predictions of future realized values of random variables based on currently available data is a frequent task in statistical applications. In some applications, the interest is to obtain a two-sided simultaneous prediction interval (SPI) to contain at least k out of m future observations with a certain confidence level based on n previous observations from the same distribution. A closely related problem is to obtain a one-sided upper (or lower) simultaneous prediction bound (SPB) to exceed (or be exceeded) by at least k out of m future observations. In this paper, we provide a general approach for constructing SPIs and SPBs based on data from a member of the (log)-location-scale family of distributions with complete or right censored data. The proposed simulation-based procedure can provide exact coverage probability for complete and Type II censored data. For Type I censored data, the simulation results show that our procedure provides satisfactory results in small samples. We use three applications to illustrate the proposed simultaneous prediction intervals and bounds.

Key Words: Censored Data; Coverage Probability; k out of m, Lognormal; Simulation; Weibull.

1 Introduction

1.1 Motivation

Prediction intervals are used to quantify the uncertainty associated with future realized values of random variables. In predicting future outcomes, one might be interested in point predictions. Often, however, the focus is on whether the future observations will fall within a prediction interval (PI) or conforming to a one-sided prediction bound (PB) obtained from the available data and a pre-specified confidence level.

In some applications, it is desirable to obtain a two-sided simultaneous prediction interval (SPI) or a one-sided simultaneous prediction bound (SPB) for at least k out of m future observations, where $1 \leqslant k \leqslant m$. For example, Fertig and Mann (1977) consider time to failure of turbine nozzles subject to a certain load. The company had manufactured 50 nozzles. Based on the failure times in a life test of 10 of those nozzles, they obtained a 95% lower prediction bound to be exceeded by at least 90% of the remaining 40 nozzles (i.e., 36 out of 40). In another study, Fertig and Mann (1977) use failure times (in hours) based on a life test of aircraft components to obtain an SPI to contain the failure times of all 10 future components.

Much research has been done for statistical prediction for a single future observation. Details and additional references can be found in Mee and Kushary (1994) and Escobar and Meeker (1999). There has been some work for the SPIs/SPBs for at least k out of m future observations. Those procedures, however, have been developed only for specific distributions (e.g., normal and Weibull distributions). Hence, it is desirable to have a general approach to generate SPIs/SPBs for a general class of distributions. In this paper, we develop a general procedure to obtain SPIs and SPBs for the location-scale family and the log-location-scale family of distributions. The proposed procedures can be used with complete or censored data and can be extended, in an approximate manner, to other distributions.

1.2 Literature Review and Contributions of This Work

There is some previous work on the construction of SPIs/SPBs to contain/bound at least k out of m future observations. Danziger and Davis (1964) described and provided tables of coverage probabilities for non-parametric SPIs to contain k out of m future observations (which they refer to as tolerance intervals) and corresponding one-sided SPBs. Hahn (1969) considered the special case of $k = m$ based on observations from a normal distribution. Hahn (1969) gave the factors to calculate two-sided SPIs. One-sided SPBs were considered in Hahn (1970). Fertig and Mann (1977) presented factors for constructing one-sided SPBs to contain

at least k out of m future observations for a normal distribution. Odeh (1990) provided a method for generating k out of m two-sided SPIs for a normal distribution. Due to computational limitations, these papers only provided factors for a limited number of combinations of n, k, m and for some specified confidence levels. In the area of environmental monitoring, some articles considered the use of SPIs/SPBs for at least k out of m future observations at p locations. Davis and McNichols (1987) studied this type of problem for one-sided prediction bounds and for observations from a normal distribution. Krishnamoorthy, Lin, and Xia (2009) constructed one-sided upper prediction bounds for the Weibull distribution based on generalized pivotal quantities. Bhaumik and Gibbons (2006) developed an approximate upper SPB for samples from a gamma distribution. Bhaumik (2008) constructed a one-sided SPB for left-censored normal random variables. Beran (1990) gives theoretical results on the coverage properties of the prediction regions based on simulation. There are no methods in the literature for two-sided SPIs for the Weibull distribution.

None of existing literature proposes a general procedure for the location-scale (e.g., the smallest extreme value, normal, and largest extreme value distributions) or the related log-location-scale family of distributions (e.g., the Weibull, lognormal, and Fréchet distributions). In this paper, we develop methods for constructing such intervals/bounds based on a general procedure. The methods are exact (except for Monte Carlo error) for complete and Type II censored data. Type I censoring is commonly in life tests. We use simulation to study the coverage properties for the approximate intervals/bounds under Type I censoring.

1.3 Overview

The rest of this paper is organized as follows. Section 2 introduces the data and model setting for the problem. Section 3 gives the formal definition of the proposed SPI procedure. Section 4 proposes a general procedure to obtain an SPI, followed by illustrative examples. Section 5 describes simulation studies on the performance of the proposed procedure for Type I censored data. Section 6 illustrates the use of the proposed method with applications. Section 7 contains concluding remarks and some discussion about related extensions and applications of the methods.

2 Data, Model, and Maximum Likelihood Estimation

2.1 Data

We consider situations in which n independent experimental units are under study. At the moment of doing the analysis, the data consist of: (a) r exact observations and (b) a set of $(n-r)$ right-censored observations at x_c, where x_c is larger or equal to the maximum of the exact observations. Three important special cases of these data structure are: (a) complete data, when $r = n$; (b) Type II censored data, when r $(2 \leqslant r \leqslant n)$ is pre-specified and x_c is equal to the maximum of the exact observations. Note that in this case x_c is random; (c) Type I censored data, when x_c is pre-specified and x_c exceeds the maximum of the exact observations. Note that in the case of Type I censoring, r $(1 \leqslant r \leqslant n)$ is random (if $r = 0$ the maximum likelihood (ML) estimate does not exist).

To be precise, let $\boldsymbol{X} = (X_1, \ldots, X_n)$ denote the random variables for the observations from the n units, where $-\infty < X_i < \infty, i = 1, \ldots, n$. Define

$$\delta_i = \begin{cases} 1, & \text{if } X_i \text{ is an exact observation} \\ 0, & \text{if } X_i \text{ is a right-censored observation} \end{cases}$$

For Type I and Type II censoring, we observe $x_i = \min(X_i, x_c)$ and δ_i, $i = 1, \ldots, n$. The observed values are denoted by $\boldsymbol{x} = (x_1, \ldots, x_n)$. This data structure is general and includes data from reliability and lifetime studies with right-censored data from a positive response. In this case all the components of $\boldsymbol{X}$ take positive values.

2.2 Model

To construct an SPI for a set of future observations, we use a statistical model to describe the population of interest. In this paper, we assume the observations have a distribution in the family of the location-scale or log-location-scale family of distributions. A location-scale distribution has a location parameter μ and a scale parameter σ. The parameters μ and σ are typically unknown and need to be estimated. The probability density function (pdf) and the cumulative distribution function (cdf) of a location-scale distribution are

$$f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) \quad \text{and} \quad F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

respectively. The definitions of the standard pdf $\phi(\cdot)$ and cdf $\Phi(\cdot)$ functions for the different members of this family are given in Table 1.

Table 1: The pdfs and cdfs of different commonly-used members of the standard location-scale and log-location-scale distributions.

Location-Scale	Log-Location-Scale	pdf $\phi(x)$	cdf $\Phi(x)$
Normal	Lognormal	$\frac{\exp(-x^2/2)}{\sqrt{2\pi}}$	$\int_{-\infty}^{x} \phi(w)\, dw$
Logistic	Loglogistic	$\frac{\exp(x)}{[1+\exp(x)]^2}$	$\frac{\exp(x)}{1+\exp(x)}$
Largest extreme value	Fréchet	$\exp[-x-\exp(-x)]$	$\exp[-\exp(-x)]$
Smallest extreme value	Weibull	$\exp[x-\exp(x)]$	$1-\exp[-\exp(x)]$

The pdf and cdf of the log-location-scale family are

$$f(t) = \frac{1}{\sigma t}\phi\left[\frac{\log(t)-\mu}{\sigma}\right] \quad \text{and} \quad F(t) = \Phi\left[\frac{\log(t)-\mu}{\sigma}\right],$$

respectively. The Weibull, lognormal, Fréchet, and log-logistic distributions are members of the log-location-scale family. For these distributions, σ is a shape parameter and $\exp(\mu)$ is a scale parameter. In the reminder of this paper, however, we will refer to μ and σ as location and scale parameters, respectively.

This paper focuses on the construction of SPIs and SPBs containing at least k of m future observations $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ from a previously sampled population. The sample data are denoted by $\boldsymbol{X}$ and the assumptions are that $\boldsymbol{Y}$ and $\boldsymbol{X}$ are independent and random samples from the same distribution.

2.3 Maximum Likelihood Estimation

We use maximum likelihood (ML) to estimate the unknown parameters (μ, σ). Under the i.i.d. assumptions in Section 2.2, the likelihood of the right censored data has the form

$$L(\mu,\sigma) = \mathcal{C}\prod_{i=1}^{n}[f(x_i;\mu,\sigma)]^{\delta_i}[1-F(x_i;\mu,\sigma)]^{1-\delta_i},$$

where $\mathcal{C}$ is a constant that does not depend on μ or σ, $f(x_i;\mu,\sigma)$ is the assumed pdf, and $F(x_i;\mu,\sigma)$ is the corresponding cdf. The ML estimates can be obtained by finding the values of μ and σ that maximize the likelihood function. In general, there is no closed-form expression for the ML estimates, which are denoted by $(\widehat{\mu}, \widehat{\sigma})$. Consequently, numerical methods are used to find the ML estimates.

3 Simultaneous Prediction Intervals and Bounds

3.1 Two-sided Simultaneous Prediction Intervals

This section shows how to construct an SPI $[L(\boldsymbol{x}, 1-\alpha),\ U(\boldsymbol{x}, 1-\alpha)]$ that will contain at least k out of m independent future observations from the sampled distribution, with a specified confidence level $1-\alpha$. Conditioning on the observed data $\boldsymbol{X} = \boldsymbol{x}$, the conditional coverage probability (CP) of the interval $[L(\boldsymbol{x}, 1-\alpha),\ U(\boldsymbol{x}, 1-\alpha)]$ with nominal confidence level $1-\alpha$ is

$$\begin{aligned}\mathrm{CP}(\boldsymbol{\theta}|\boldsymbol{X}=\boldsymbol{x}) &= \Pr\{\text{at least } k \text{ of } m \text{ values lie in } [L(\boldsymbol{x}, 1-\alpha),\ U(\boldsymbol{x}, 1-\alpha)]\,|\boldsymbol{X}=\boldsymbol{x}\} \\ &= \sum_{j=k}^{m} \binom{m}{j} p^j (1-p)^{m-j}, \end{aligned} \tag{1}$$

where $\boldsymbol{\theta} = (\mu, \sigma)$ is the vector of unknown parameters and

$$p = \Pr\{\text{a future observation is in } [L(\boldsymbol{x}, 1-\alpha),\ U(\boldsymbol{x}, 1-\alpha)]\,|\boldsymbol{X}=\boldsymbol{x}\}.$$

The conditional CP is unobservable because it depends on the unknown parameters and varies from sample to sample because it depends on the data. Following standard procedure, to evaluate the prediction interval procedure, we use the unconditional CP

$$\mathrm{CP}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{X}}\left[\sum_{j=k}^{m} \binom{m}{j} p^j (1-p)^{m-j}\right],$$

where expectation is taken with respect to the joint distribution of the data $\boldsymbol{X}$.

Because $(Y_i - \widehat{\mu})/\widehat{\sigma}, i = 1, \ldots, m$ are pivotal quantities, one can construct a two-sided $100(1-\alpha)\%$ SPI to contain at least k out of m future observations with the following form

$$[\,\widehat{\mu} + u_L(k, m; \alpha)\widehat{\sigma},\ \widehat{\mu} + u_U(k, m; \alpha)\widehat{\sigma}\,],$$

for the location-scale family of distributions. Here $u_L(k, m; \alpha)$ and $u_U(k, m; \alpha)$ are factors to be chosen so that the SPI will have CP equal to $1-\alpha$. For notational simplicity, we let $u_L = u_L(k, m; \alpha)$ and $u_U = u_U(k, m; \alpha)$. In particular, the factors (u_L, u_U) satisfy the equation

$$1-\alpha = \int_0^\infty \int_{-\infty}^\infty \sum_{j=k}^{m} \binom{m}{j} [\Phi(a) - \Phi(b)]^j\, [1 - \Phi(a) + \Phi(b)]^{m-j}\, f_{\boldsymbol{Z}}(z_1, z_2)\, dz_1 dz_2, \tag{2}$$

where $a = z_1 + u_U z_2$, $b = z_1 + u_L z_2$, $\boldsymbol{Z} = (Z_1, Z_2)$, $f_{\boldsymbol{Z}}(z_1, z_2)$ is the joint pdf of $Z_1 = (\widehat{\mu} - \mu)/\sigma$ and $Z_2 = \widehat{\sigma}/\sigma$, and $\Phi(\cdot)$ is the standard cdf of X.

The proof of (2) is given in Appendix A. Note that (2) can be written as

$$1-\alpha = \mathbf{E}_{\boldsymbol{Z}}\left[\sum_{j=k}^{m}\binom{m}{j}\left[\Phi(A)-\Phi(B)\right]^{j}\left[1-\Phi(A)+\Phi(B)\right]^{m-j}\right], \tag{3}$$

where $A = Z_1 + u_U Z_2$, $B = Z_1 + u_L Z_2$, and $\mathbf{E}_{\boldsymbol{Z}}(\cdot)$ is the expectation with respect to the joint distribution of $\boldsymbol{Z}$.

For distributions in the log-location-scale family, the corresponding two-sided $100(1-\alpha)\%$ SPI to contain at least k out of m future observations has the form $[\exp(\widehat{\mu}+u_L\widehat{\sigma}), \exp(\widehat{\mu}+u_U\widehat{\sigma})]$. Thus, (2) is still used to obtain a prediction interval for distributions in the log-location-scale family.

For Type II censored data or complete data from the location-scale/log-location-scale family of distributions, Lawless (2003, pages 217 and 262) describes the pivotal property of $\boldsymbol{Z}$. That is, the distribution of $\boldsymbol{Z}$ does not depend on unknown parameters. For Type I censoring, the pivotal property of $\boldsymbol{Z}$ no longer holds. The quantity $\boldsymbol{Z}$, however, can be treated as being approximately pivotal. Thus we can still use (2) to get the approximate asymptotically correct SPIs under Type I censoring, and other types of non-informative censoring.

3.2 One-sided Simultaneous Prediction Bounds

There are similar CP statements for one-sided simultaneous prediction bounds. In particular, for a one-sided lower simultaneous prediction bound, the conditional CP is

$$\begin{aligned}\mathrm{CP}_L(\boldsymbol{\theta}|\boldsymbol{X}=\boldsymbol{x}) &= \Pr\left[\text{at least } k \text{ of } m \text{ values are larger than } L(\boldsymbol{x},1-\alpha)|\boldsymbol{X}=\boldsymbol{x}\right]\\ &= \sum_{j=k}^{m}\binom{m}{j}p^{j}(1-p)^{m-j},\end{aligned} \tag{4}$$

where $p = \Pr\left[\text{a single future observation is larger than } L(\boldsymbol{x},1-\alpha)|\boldsymbol{X}=\boldsymbol{x}\right]$.

The unconditional CP is

$$\mathrm{CP}_L(\boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{X}}\left[\sum_{j=k}^{m}\binom{m}{j}p^{j}(1-p)^{m-j}\right].$$

For the location-scale family of distributions, a one-sided lower simultaneous prediction bound to be exceeded by at least k out of m future observations can be expressed as $L(\boldsymbol{x},1-\alpha) = \widehat{\mu}+u'_L(k,m;\alpha)\widehat{\sigma}$, where $u'_L(k,m;\alpha)$ is a factor to be chosen so that the

interval will give a CP of $1-\alpha$. Let $u'_L = u'_L(k, m; \alpha)$ and note that u'_L satisfies the equation

$$\begin{aligned} 1-\alpha &= \int_0^\infty \int_{-\infty}^\infty \sum_{j=k}^m \binom{m}{j} \left[1-\Phi(b)\right]^j \left[\Phi(b)\right]^{m-j} \, f_{\boldsymbol{Z}}(z_1, z_2)\, dz_1 dz_2 \\ &= \mathbf{E}_{\boldsymbol{Z}} \left[\sum_{j=k}^m \binom{m}{j} \left[1-\Phi(B)\right]^j \left[\Phi(B)\right]^{m-j} \right], \end{aligned} \tag{5}$$

where $b = z_1 + u'_L z_2$ and $B = Z_1 + u'_L Z_2$. When $k = m$, one obtains the lower prediction bound to contain all m new additional observations.

Similarly, for a one-sided upper simultaneous prediction bound, the conditional CP is

$$\begin{aligned} \mathrm{CP}_U(\boldsymbol{\theta}|\boldsymbol{X} = \boldsymbol{x}) &= \Pr\left[\text{at least } k \text{ of } m \text{ values are less than } L(\boldsymbol{x}, 1-\alpha)|\boldsymbol{X} = \boldsymbol{x}\right] \\ &= \sum_{j=k}^m \binom{m}{j} p^j (1-p)^{m-j}, \end{aligned} \tag{6}$$

where

$$p = \Pr\left[\text{a single future observation is less than } L(\boldsymbol{x}, 1-\alpha)|\boldsymbol{X} = \boldsymbol{x}\right].$$

The unconditional CP is

$$\mathrm{CP}_U(\boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{X}} \left[\sum_{j=k}^m \binom{m}{j} p^j (1-p)^{m-j} \right].$$

A one-sided upper simultaneous prediction bound to exceed at least k out of m future observations for the location-scale family of distributions is $U(\boldsymbol{x}, 1-\alpha) = \widehat{\mu} + u'_U(k, m; \alpha)\widehat{\sigma}$, where $u'_U(k, m; \alpha)$ is a factor to be chosen so that the interval will give a CP equal to $1-\alpha$. Let $u'_U = u'_U(k, m; \alpha)$ and note that u'_U satisfies the equation

$$\begin{aligned} 1-\alpha &= \int_0^\infty \int_{-\infty}^\infty \sum_{j=k}^m \binom{m}{j} \left[\Phi(a)\right]^j \left[1-\Phi(a)\right]^{m-j} \, f_{\boldsymbol{Z}}(z_1, z_2)\, dz_1 dz_2 \\ &= \mathbf{E}_{\boldsymbol{Z}} \left[\sum_{j=k}^m \binom{m}{j} \left[\Phi(A)\right]^j \left[1-\Phi(A)\right]^{m-j} \right], \end{aligned} \tag{7}$$

where $a = z_1 + u'_U z_2$ and $A = Z_1 + u'_U Z_2$.

For the log-location-scale family of distributions, the lower and upper SPBs have the form $L(\boldsymbol{x}, 1-\alpha) = \exp(\widehat{\mu} + u'_L\widehat{\sigma})$ and $U(\boldsymbol{x}, 1-\alpha) = \exp(\widehat{\mu} + u'_U\widehat{\sigma})$, respectively. The factors u'_L and u'_U are obtained as solutions of (5) and (7), respectively.

4 Computations of the Simultaneous Prediction Intervals/Bounds

In this section, we introduce a general procedure for finding the factors so that the two-sided SPIs and one-sided SPBs will have the correct CP. The computing procedure requires solving equations (3), (5), and (7). In general, there is no closed-form expression for the solution of these equations. The exact distribution of $\boldsymbol{Z}$ can be complicated, especially with censored data. Therefore, we use Monte Carlo simulation to obtain the distribution of $\boldsymbol{Z}$ and evaluate the expectation based on the simulated samples.

4.1 Complete and Type II Censored Data

The two-sided SPI for complete or Type II censored data can be obtained from the following algorithm.

Algorithm 1:

1. Draw a complete or Type II censored sample of size n from a (log)-location-scale family of distributions with $(\mu, \sigma) = (0, 1)$. Detailed discussion on efficient simulation of censored samples can be found in Meeker and Escobar (1998, Section 4.13).

2. Repeat step 1 B_1 times and compute ML estimates $(\widehat{\mu}_l^*, \widehat{\sigma}_l^*)$ for each simulated sample, $l = 1, \ldots, B_1$.

   To save computing time, these $(\widehat{\mu}_l^*, \widehat{\sigma}_l^*)$ values are stored and used to compute all the SPIs and SPBs for the particular censoring specification (n, r) as shown below.

3. For every (u_L, u_U), in a collection of chosen values, compute

$$\mathrm{CP}^*(u_L, u_U) = \frac{1}{B_1} \sum_{l=1}^{B_1} \left\{ \sum_{j=k}^{m} \binom{m}{j} p_l(u_L, u_U)^j [1 - p_l(u_L, u_U)]^{m-j} \right\}, \tag{8}$$

   where $p_l(u_L, u_U) = \Phi(\widehat{\mu}_l^* + u_U \widehat{\sigma}_l^*) - \Phi(\widehat{\mu}_l^* + u_L \widehat{\sigma}_l^*)$ and $u_L < u_U$.

4. Find (u_L, u_U) such that $\mathrm{CP}^*(u_L, u_U) = 1 - \alpha$.

Note that the choice of $(\mu, \sigma) = (0, 1)$ in Step 1 above is justified because for the Type II censored and complete data case, the **Algorithm 1** procedure does not depend on unknown parameters due to the pivotal property of $\boldsymbol{Z}$.

Finding (u_L, u_U) such that $\mathrm{CP}^*(u_L, u_U) = 1 - \alpha$ is a two-dimensional root-finding problem and there are multiple solutions. An additional constraint on u_L and u_U is needed for a

unique solution. For symmetric distributions, $u_L = -u_U$ is an appropriate constraint and leads to two-sided SPIs with equal tail probabilities. For non-symmetric distributions, the two-sided SPI with equal tail probabilities is appealing from a practical point of view. The computation, however, is more complicated. Detailed discussion of the computation is given in Section 4.2.

For one-sided SPBs, modifications to the algorithm are needed. Specifically, for the lower SPB, replace (8) by

$$\mathrm{CP}_L^*(u'_L) = \frac{1}{B_1}\sum_{l=1}^{B_1}\left\{\sum_{j=k}^{m}\binom{m}{j}p_l(u'_L)^j[1-p_l(u'_L)]^{m-j}\right\},$$

where $p_l(u'_L) = 1-\Phi(\widehat{\mu}_l^* + u'_L\widehat{\sigma}_l^*)$. Then find the unique value of u'_L such that $\mathrm{CP}_L^*(u'_L) = 1-\alpha$. For the upper SPB, we need to replace (8) by

$$\mathrm{CP}_U^*(u'_U) = \frac{1}{B_1}\sum_{l=1}^{B_1}\left\{\sum_{j=k}^{m}\binom{m}{j}p_l(u'_U)^j[1-p_l(u'_U)]^{m-j}\right\},$$

where $p_l(u'_U) = \Phi(\widehat{\mu}_l^* + u'_U\widehat{\sigma}_l^*)$. Then find the unique value of u'_U such that $\mathrm{CP}_U^*(u'_U) = 1-\alpha$. For one-sided prediction bounds, we use linear interpolation to obtain lower or upper limits based on the CP curve ($1-\alpha$ versus u'_L or u'_U, respectively) for desired confidence levels.

4.2 Two-sided SPI with Equal Tail Probability

In applications, even involving a non-symmetric distribution, it is preferable to have a two-sided prediction interval with equal tail probabilities. For this purpose, we define the tail probability as the tail probability of the one-sided bound. Therefore, the equal tail probability implies that $\mathrm{CP}_L(u_L) = \mathrm{CP}_U(u_U)$. Except for the special case of $k = 1$ (i.e., a prediction interval for exactly one new observation), combining a one-sided lower $100(1-\alpha_1)\%$ prediction bound and a one-sided upper $100(1-\alpha_2)\%$ prediction bound will not provide a two-sided $100(1-\alpha_1-\alpha_2)\%$ SPI. Thus, a special procedure for a two-sided SPI with equal tail probabilities is needed. For a given confidence level $1-\alpha$, we can obtain u_L and u_U by solving numerically the equations

$$\mathrm{CP}(u_L, u_U) = 1-\alpha \quad \text{and} \quad \mathrm{CP}_L(u_L) - \mathrm{CP}_U(u_U) = 0. \tag{9}$$

To find the solutions to (9), one finds numerically the contour lines of $\mathrm{CP}(u_L, u_U)$ and $\mathrm{CP}_L(u_L) - \mathrm{CP}_U(u_U)$. Then use interpolation to locate the intersecting point of the contours. It is also possible to re-express the two-sided CP as a function of the one-sided tail probability to reduce the dimension of root-finding, and then find the common tail probability that gives the desired two-sided CP. Illustration of this method is given in Section 4.4.

4.3 Type I Censored Data

For Type I censored data, the statistics Z_1 and Z_2 are only approximately pivotal. The simulation procedure will depend on the censoring time (or more precisely, the estimated expected fraction failing). Thus for Type I censoring, we use the following algorithm.

Algorithm 2:

1. For the observed Type I data, calculate the ML estimates $(\widehat{\mu}, \widehat{\sigma})$.
2. Draw a censored sample of size n from the (log)-location-scale family of distributions with $(\mu, \sigma) = (\widehat{\mu}, \widehat{\sigma})$ and censoring mimicking the censoring in the original data.
3. Repeat step 2 B_1 times and compute ML estimates $(\widehat{\mu}_l^*, \widehat{\sigma}_l^*)$ for each simulated sample, $l = 1, \ldots, B_1$.
4. For every (u_L, u_U), in a collection of chosen values, compute

$$\mathrm{CP}^*(u_L, u_U) = \frac{1}{B_1} \sum_{l=1}^{B_1} \left\{ \sum_{j=k}^{m} \binom{m}{j} p_l(u_L, u_U)^j [1 - p_l(u_L, u_U)]^{m-j} \right\}, \tag{10}$$

   where $p_l(u_L, u_U) = \Phi[(\widehat{\mu}_l^* + u_U \widehat{\sigma}_l^* - \widehat{\mu})/\widehat{\sigma})] - \Phi[(\widehat{\mu}_l^* + u_L \widehat{\sigma}_l^* - \widehat{\mu})/\widehat{\sigma}]$ and $u_L < u_U$.
5. Find (u_L, u_U) such that $\mathrm{CP}^*(u_L, u_U) = 1 - \alpha$.

As the sample size increases, the CP of the SPIs/SPBs for Type I censoring data computed by **Algorithm 2** will approach the nominal confidence level. In Section 5, we study finite sample CPs for SPIs and SPBs obtained using **Algorithm 2**.

4.4 Illustrative Examples

Illustration A: Upper SPB for Type II Censoring and Complete Data

For purpose of illustration, we generate the CP curve for a one-sided upper SPB for at least 4 out of 5 future observations from a previous sampled Weibull distribution. The sample size is $n = 20$ and we consider the Type II censored configurations corresponding to $r = 5, 10, 15,$ and 20 (complete data case). The number of simulations B_1 is set to be 100,000 so that the results are stable (i.e., negligible Monte Carlo error). Figure 1 shows the CP as a function of u'_U and r. For a desired coverage level, say $1 - \alpha = 0.95$ and a specific value of r, the value of u'_U is determined from the CP curve corresponding to the specified r value.

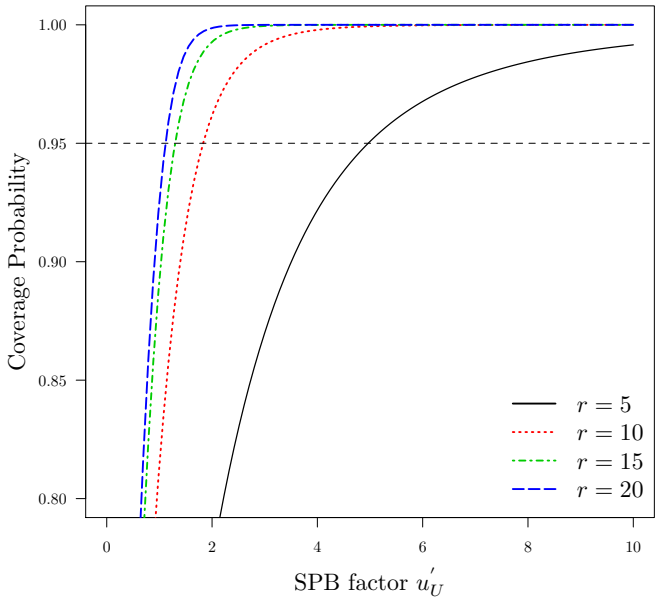


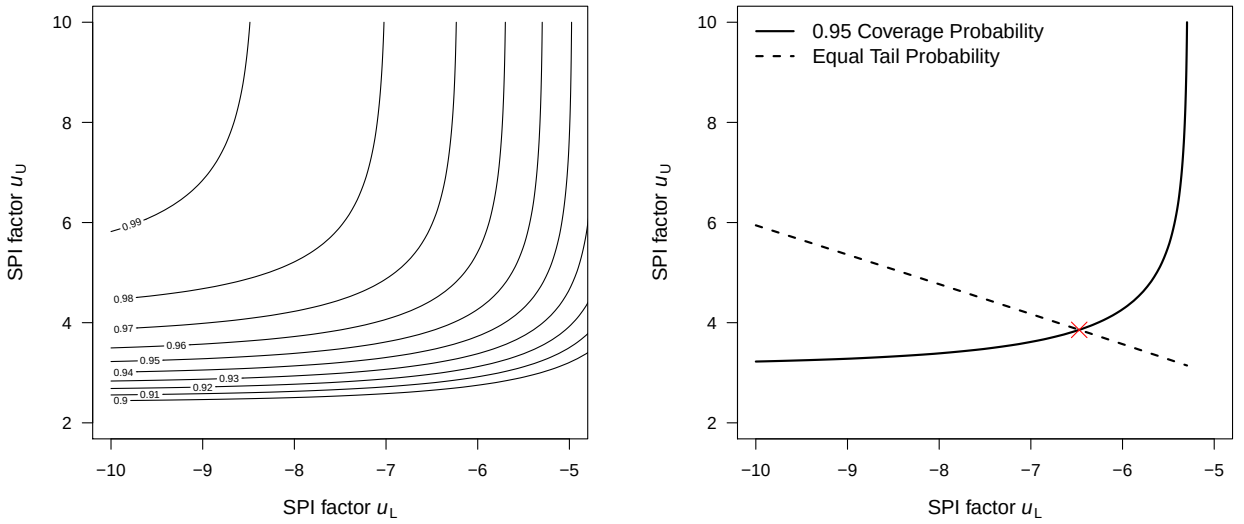
Figure 1: CP curves of one-sided upper SPBs for $n = 20, r = 5, 10, 15$ and $20, k = 4,$ and $m = 5$ based on **Algorithm 1**.

Illustration B: Two-sided SPI with Equal Tail Probability

Here the objective is to construct a two-sided SPI with equal tail probabilities from a previously sampled Weibull distribution. Again, B_1 is chosen to be 100,000. The contour plot of the CP as a function of u_L and u_U is shown in Figure 2(a). To obtain the prediction interval with equal probability in each tail, we solve the equations in (9). Figure 2(b) shows the contour lines of the two equations. The upper and lower limits u_U and u_L of the 95% SPI with equal tail probabilities are the coordinates of the intersection point of the two non-linear curves in Figure 2(b). In Figure 2(b), the coordinates $(u_L, u_U) = (-6.46, 3.82)$ produce both the 0.95 overall coverage probability and the equal tail probabilities.

5 Simulation Study for Type I censoring

This section studies the CP properties of the simulation-based procedure proposed in Section 4.3. For the Type I censored data case, the procedure properties will depend on unknown parameters through the censoring time (or, more generally, the expected fraction failing). The CP of the SPIs/SPBs, however, will converge to the nominal confidence level as the ex-



(a) Coverage Probability Plot

(b) 0.95 Coverage and Equal Tail Probability Plot

Figure 2: (a) Contour plot of the CP as a function of u_L and u_U for $n = 20, r = 8, k = m = 3$, based on 100,000 simulations. (b)The contour lines of equation (9).

pected number failing increases to infinity. Here we study the effect of the expected number of failures r_f on the CP of the SPIs/SPBs in small samples. Similar simulation designs can be found in Vander Weil and Meeker (1990) and Jeng and Meeker (2001). In **Algorithm 2**, we calculated the SPIs/SPBs based on the ML estimates $(\widehat{\mu}, \widehat{\sigma})$, which are determined from the observed data. To evaluate the performance of **Algorithm 2**, we simulate the data many times and average over the results. The detailed simulation plan is as follows.

1. Simulate $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ with the pre-determined censoring time. Without loss of generality, we simulate samples from the Weibull distribution with parameters $(\mu, \sigma) = (0, 1)$. Then, calculate the ML estimates of (μ, σ) for each simulated sample.

2. Use **Algorithm 2** to obtain the SPIs/SPBs. For example, we can obtain the one-sided upper SPB by computing u'_U.

3. Use (1), (4), and (6) to compute the conditional CP for the SPI, the lower SPB, and the upper SPB, respectively.

4. Repeat the steps $1 - 3$ B_2 times and obtain the estimates of the unconditional CP for the SPIs/SPBs by averaging over the conditional CPs.

Because the focus is on the CP for small sample sizes, we simulate datasets with the expected number of failures equal to $r_f = 5, 7, 10, 25$ and the expected fraction failing equal to $p_f = 0.25$. Here we chose $B_2 = 500$ for the purpose of controlling the computational cost while maintaining a reasonably small Monte Carlo error.

Figure 3 displays the estimated actual CP versus the nominal confidence level for the one-sided lower and upper SPBs, and the two-sided SPI. Figure 3 shows that there are some deviations from the nominal CP when the expected number of failures r_f is small (around 10). The estimated actual CP is close to the nominal confidence level when r_f is large enough (e.g., around 25). In the case of $r_f = 25$, the corresponding line is nearly the same as the identity line. When r_f is large, the observed data tends to have more failures, thus the estimates are more accurate and the SPIs/SPBs have better CP. We also note that the two-sided SPI tends to perform better than one-sided SPBs when r_f is small. As indicated earlier, we used $(\mu, \sigma) = (0, 1)$ in the simulation. For other values of (μ, σ), the simulation results are similar because they depend on the expected number of failures. Overall, **Algorithm 2** provides satisfactory results for Type I censoring in finite samples when the expected number of failures is at least 5.

6 Applications

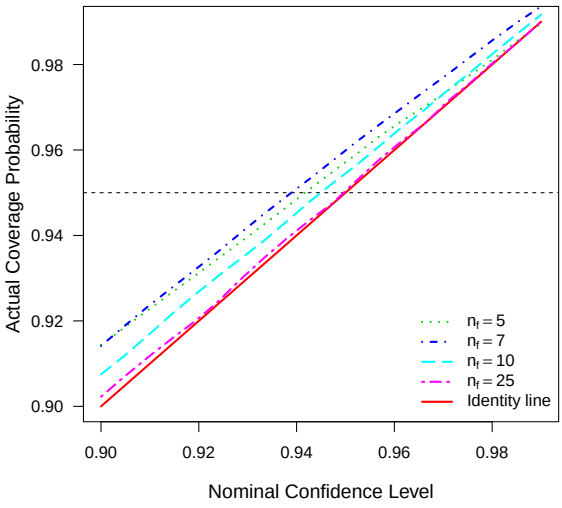
In this section, we use three examples to illustrate the applicability of the proposed procedure.

6.1 Nozzle Failure Time Data

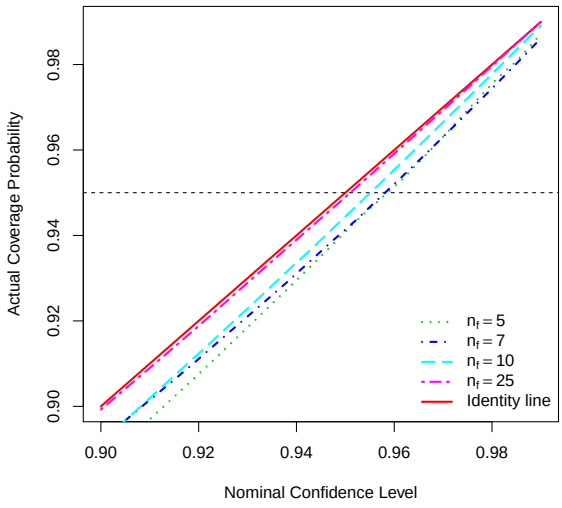
This example is adopted from the application described in Fertig and Mann (1977). They wanted to compute a 95% lower prediction bound (they called a "warranty period") of the failure times of at least 36 or 40 out of 40 nozzles. They provided the sample mean and sample standard derivation of the logarithm of failure times (which they assumed to have normal distribution) of 10 nozzles, which are $\widehat{\mu} = 3.850$ and $\widehat{\sigma} = 0.034$, respectively. Applying **Algorithm 1**, we found that the lower SPBs for at least 36 and at least 40 out of 40 nozzles to be 43.35 and 40.96 hours (based on 100,000 Monte Carlo trials), respectively.

6.2 Aircraft Component Failure Time Data

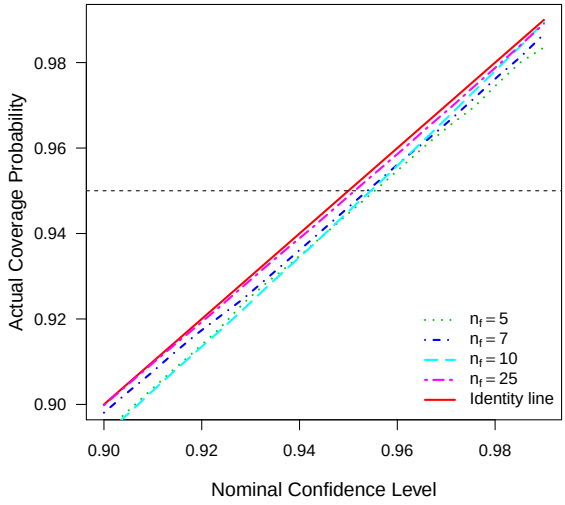
Mann and Fertig (1973) describes a study yielding ten failure times out of 13 aircraft components that were tested. The failure times were 0.22, 0.50, 0.88, 1.00, 1.32, 1.33, 1.54, 1.76, 2.50, and 3.00 hours. The three right censored observations occurred at 3.00 hours. Both



(a) Lower SPB



(b) Upper SPB



(c) Two-sided SPI

Figure 3: Estimated actual CP versus nominal confidence level for fixed $p_f = 0.25$, when $k = 4$ and $m = 5$. (a) Lower SPB. (b) Upper SPB. (c) Two-sided SPI.

Mann and Fertig (1973) and Hsieh (1996) state that it is reasonable to assume a Weibull model for the data. The Weibull probability plot in Figure 4(a) corroborates the adequacy of the Weibull model. Based on Figure 4(b), the lognormal distribution, however, is also suitable to describe the failure-time distribution of the aircraft component. Using **Algorithm 1** one obtains 95% lower SPBs of the failure times of all 10 future aircraft components, which are 0.003 hours and 0.04 hours for the Weibull and lognormal distributions, respectively. Also we found that the 95% upper SPBs are 39.789 hours and 107.465 hours for the Weibull and lognormal distributions, respectively. The large difference is due to the implied extrapolation, especially into the upper tail of the failure-time distribution.

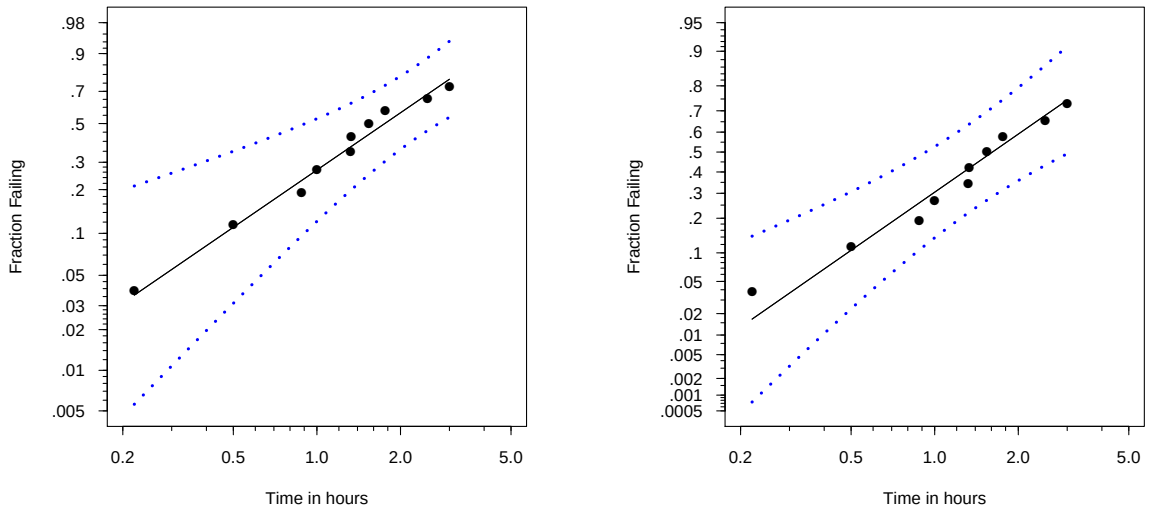
6.3 Vinyl Chloride Data

This application uses data consisting of vinyl chloride concentrations (in $\mu g/L$) from clean upgradient ground-water monitoring wells. The data were given in Bhaumik and Gibbons (2006). The probability plot in Bhaumik and Gibbons (2006) indicates that the gamma distribution fits the data well. Figures 5(a) and 5(b) indicate that the Weibull and lognormal distributions also provide good fit to the vinyl chloride data. Bhaumik and Gibbons (2006) wanted to obtain a 95% upper SPB to exceed at least $k = 1$ out of $m = 2$ future observations. For the gamma distribution, the 95% upper SPB is 2.931 $\mu g/L$. Using **Algorithm 1**, for the Weibull distribution, the 95% upper SPB is $\exp(0.635 + 0.464 \times 0.99) = 2.989$ $\mu g/L$; for the lognormal distribution, the 95% upper SPB is $\exp(0.092 + 0.829 \times 1.120) = 2.773$ $\mu g/L$. For this application, the 95% upper SPBs for the gamma, Weibull, and lognormal distributions are closed to each other. This is because extrapolation is not required to construct this interval.

7 Concluding Remarks and Areas for Future Research

In this paper, we propose a general method for constructing simultaneous two-sided prediction intervals for at least k out of m future observations as well as the corresponding one-sided bounds for the (log)-location-scale family of distributions. For the Type II censored or complete data cases, the method provides a procedure with CP equal to the nominal confidence level (ignoring Monte Carlo error that can be made arbitrarily small). For Type I censored data, the approximate procedure provides coverage probabilities that are close to the nominal confidence level if the expected number of failures is not too small.

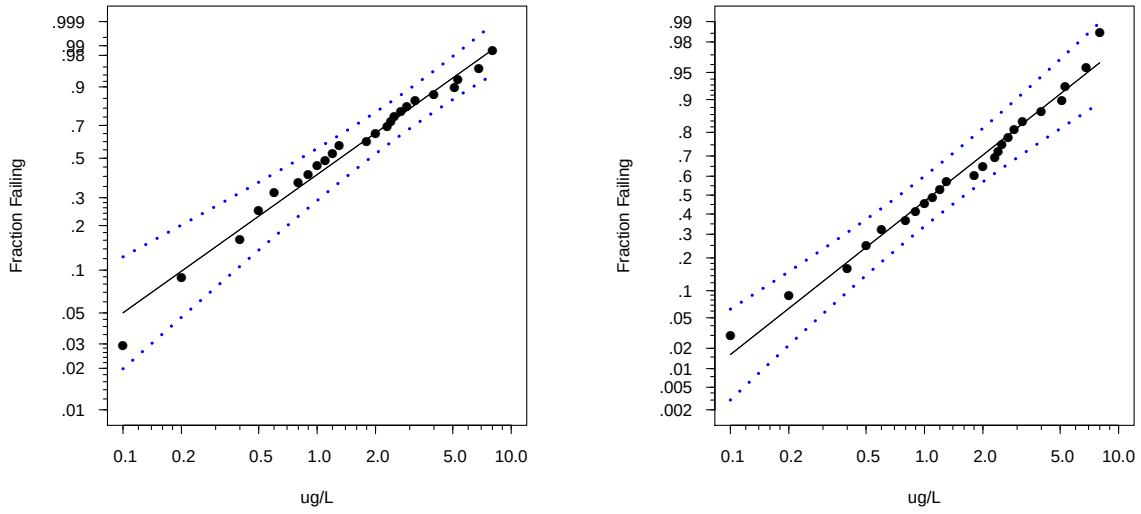
The procedures in this paper can also be extended to data involving multiple censoring or random censoring. With complete data, the extension of the proposed methods to regression



(a) Weibull

(b) lognormal

Figure 4: Probability plots for the aircraft data. (a) Weibull fit. (b) Lognormal fit.



(a) Weibull

(b) lognormal

Figure 5: Probability plots for the vinyl chloride data. (a) Weibull fit. (b) Lognormal fit.

case is straightforward because the pivotal properties still hold (Lawless 2003, Appendix E4). As long as the pivotal property holds, the proposed procedure can be easily extended to give exact prediction intervals. When the pivotal property no longer holds (e.g., with regression and censoring), the approximate pivotal approach can be applied.

A Proof of Equation (2)

Let A_j be the event that exactly j of the future observations $\boldsymbol{Y}$ lie in the prediction interval $[\widehat{\mu} + u_L\widehat{\sigma},\ \widehat{\mu} + u_U\widehat{\sigma}]$. To compute $\Pr(A_j)$, first we compute the conditional probability $\Pr(A_j|\widehat{\mu},\widehat{\sigma})$ and then average this conditional probability over the sampling distribution of the ML estimates $(\widehat{\mu},\widehat{\sigma})$.

Now we proceed to compute $\Pr(A_j|\widehat{\mu},\widehat{\sigma})$. Define the indicator variables

$$I_j = \begin{cases} 1 & \text{if } Y_j \in [\widehat{\mu} + u_L\widehat{\sigma},\ \widehat{\mu} + u_U\widehat{\sigma}] \\ 0 & \text{otherwise} \end{cases}$$

where $j = 1, \ldots, m$.

The I_j variables are independent and identically distributed (iid) because the Y_j are iid. The I_j are Bernoulli(p) distributed where the p parameter is given in (11). Consequently, the number of future observations, say $S = \sum_{j=1}^{m} I_j$, contained by the conditional prediction interval $[\widehat{\mu} + u_L\widehat{\sigma},\ \widehat{\mu} + u_U\widehat{\sigma}]$ is Binomial(m, p) distributed.

The parameter p is

$$\begin{aligned} p = P(I_j = 1|\widehat{\mu},\widehat{\sigma}) &= \Pr(\widehat{\mu} + u_L\widehat{\sigma} \le Y_j \le \widehat{\mu} + u_U\widehat{\sigma}) \\ &= \Pr(Y_j \le \widehat{\mu} + u_U\widehat{\sigma}) - \Pr(Y_j \le \widehat{\mu} + u_L\widehat{\sigma}) \\ &= \Phi\left(\frac{\widehat{\mu} - \mu + u_U\widehat{\sigma}}{\sigma}\right) - \Phi\left(\frac{\widehat{\mu} - \mu + u_L\widehat{\sigma}}{\sigma}\right) \\ &= \Phi\left(\frac{\widehat{\mu} - \mu}{\sigma} + u_U\frac{\widehat{\sigma}}{\sigma}\right) - \Phi\left(\frac{\widehat{\mu} - \mu}{\sigma} + u_L\frac{\widehat{\sigma}}{\sigma}\right) \\ &= \Phi(a) - \Phi(b) \end{aligned} \tag{11}$$

where $a = z_1 + u_U z_2$, $b = z_1 + u_L z_2$, with z_1 and z_2 being realizations of the pivotals $Z_1 = (\widehat{\mu} - \mu)/\sigma$ and $Z_2 = \widehat{\sigma}/\sigma$, respectively. The value of p is the same for all the variables I_j, $j = 1, \ldots, m$, because its value does not depend on the variable Y_j chosen to do the probability computation in (11).

Thus

$$\Pr(A_j \mid \widehat{\mu},\widehat{\sigma}) = \Pr(S = j) = \binom{m}{j} p^j (1-p)^{m-j}$$

and the unconditional probability for A_j is

$$\Pr(A_j) = \int_0^\infty \int_{-\infty}^\infty \binom{m}{j} p^j (1-p)^{m-j} f_{(L,S)}(\widehat{\mu}, \widehat{\sigma})\, d\widehat{\mu} d\widehat{\sigma}$$

where $f_{(L,S)}(\widehat{\mu}, \widehat{\sigma})$ is the sampling distribution of $(\widehat{\mu}, \widehat{\sigma})$.

Define M to be the number of future observations contained by the prediction interval $[\widehat{\mu} + u_L\widehat{\sigma},\ \widehat{\mu} + u_U\widehat{\sigma}]$. Then the probability that the prediction interval contains at least k out of m future observations is

$$\begin{aligned}
\Pr(M \geq k) &= \sum_{j=k}^{m} \Pr(A_j) \\
&= \int_0^\infty \int_{-\infty}^\infty \sum_{j=k}^{m} \binom{m}{j} [\Phi(a) - \Phi(b)]^j \left[1 - \Phi(a) + \Phi(b)\right]^{m-j} f_{(L,S)}(\widehat{\mu}, \widehat{\sigma})\, d\widehat{\mu} d\widehat{\sigma} \\
&= \mathbf{E}\left[\sum_{j=k}^{m} \binom{m}{j} [\Phi(A) - \Phi(B)]^j \left[1 - \Phi(A) + \Phi(B)\right]^{m-j} \right].
\end{aligned}$$

Using (2), (u_L, u_U) can be chosen (selected/computed) to ensure that CP is equal to $(1-\alpha)$.

References

Beran, R. (1990). Calibrating prediction regions. *Journal of the American Statistical Association 85*, 715–723.

Bhaumik, D. K. (2008). One-sided simultaneous prediction limits for left-censored normal random variables. *The Indian Journal of Statistics 70-B*, 248–266.

Bhaumik, D. K. and R. D. Gibbons (2006). One-sided approximate prediction intervals for at least p of m observations from a Gamma population at each of r locations. *Technometrics 48*, 112–119.

Danziger, L. and S. A. Davis (1964). Tables of distribution-free tolerance limits. *Annals of Mathematical Statistics 35*, 1361–1365.

Davis, C. B. and R. J. McNichols (1987). One-sided intervals for at least p of m observations from a normal population on each of r future occasions. *Technometrics 29*, 359–370.

Escobar, L. A. and W. Q. Meeker (1999). Statistical prediction based on censored life data. *Technometrics 41*, 113–124.

Fertig, K. W. and N. R. Mann (1977). One-sided prediction intervals for at least p out of m future observations from a normal population. *Technometrics 19*, 167–177.

Hahn, G. J. (1969). Factors for calculating two-sided prediction intervals for samples from a normal distribution. *Journal of the American Statistical Association 65*, 878–888.

Hahn, G. J. (1970). Additional factors for calculating prediction intervals for samples from a normal distribution. *Journal of the American Statistical Association 65*, 1668–1676.

Hsieh, H. K. (1996). Prediction intervals for Weibull observations, based on early-failure data. *IEEE Transactions on Reliability 45*, 666–670.

Jeng, S.-L. and W. Q. Meeker (2001). Simultaneous parametric confidence bands for cumulative distributions from censored data. *Technometrics 43*, 450–461.

Krishnamoorthy, K., Y. Lin, and Y. Xia (2009). Confidence limits and prediction limits for a Weibull distribution based on the generalized variable approach. *Journal of Statistical Planning and Inference 139*, 2675–2684.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data, 2nd Edition*. New York: John Wiley and Sons.

Mann, N. R. and K. W. Fertig (1973). Tables for obtaining Weibull confidence bounds and tolerance bounds based on best linear invariant estimates of parameters of the extreme-value distribution. *Technometrics 15*, 87–101.

Mee, R. W. and D. Kushary (1994). Prediction limits for the Weibull distribution utilizing simulation. *Computational Statistics & Data Analysis 17*, 327–336.

Meeker, W. Q. and L. A. Escobar (1998). *Statistical methods for reliability data*. John Wiley and Sons.

Odeh, R. E. (1990). Two-sided prediction intervals to contain at least k out of m future observations from a normal distribution. *Technometrics 32*, 203–216.

Vander Weil, S. A. and W. Q. Meeker (1990). Accuracy of approx confidence bounds using censored Weibull regression data from accelerated life tests. *IEEE Transactions on Reliability 39*, 346–351.