Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Predicting the Number of Future Events

Qinglong Tian, Fanqi Meng, Daniel J. Nordman & William Q. Meeker

To cite this article: Qinglong Tian, Fanqi Meng, Daniel J. Nordman & William Q. Meeker (2021): Predicting the Number of Future Events, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1850461

To link to this article: https://doi.org/10.1080/01621459.2020.1850461

View supplementary material

Published online: 14 Jan 2021.

Submit your article to this journal

Article views: 360

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–15: Theory and Methods
https://doi.org/10.1080/01621459.2020.1850461

Taylor & Francis
Taylor & Francis Group

Check for updates

Predicting the Number of Future Events

Qinglong Tian, Fanqi Meng, Daniel J. Nordman, and William Q. Meeker

Department of Statistics, Iowa State University, Ames, IA

ABSTRACT

This article describes prediction methods for the number of future events from a population of units associated with an on-going time-to-event process. Examples include the prediction of warranty returns and the prediction of the number of future product failures that could cause serious threats to property or life. Important decisions such as whether a product recall should be mandated are often based on such predictions. Data, generally right-censored (and sometimes left truncated and right-censored), are used to estimate the parameters of a time-to-event distribution. This distribution can then be used to predict the number of events over future periods of time. Such predictions are sometimes called within-sample predictions and differ from other prediction problems considered in most of the prediction literature. This article shows that the plug-in (also known as estimative or naive) prediction method is not asymptotically correct (i.e., for large amounts of data, the coverage probability always fails to converge to the nominal confidence level). However, a commonly used prediction calibration method is shown to be asymptotically correct for within-sample predictions, and two alternative predictive-distribution-based methods that perform better than the calibration method are presented and justified. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2020
Accepted November 2020

KEYWORDS
Binomial predictand; Bootstrap; Calibration; Censored data; Predictive distribution

1. Introduction

There are many applications where it is necessary to predict the number of future events from a population of units associated with an on-going time-to-event process. Such applications also require a prediction interval to quantify statistical prediction uncertainty arising from the combination of process variability and parameter uncertainty. Some motivating applications are given below.

1.1. Product-A Data

This example is from Escobar and Meeker (1999), where, during a particular month, $n = 10{,}000$ units of Product-A were put into service. Over the next 48 months, 80 failures occurred and the failure times were recorded. A prediction interval on the number of failures among the remaining 9920 units during the next 12 months was requested by the management.

1.2. Heat Exchanger Tube Data

This example is based on data described in Nelson (2000). Nuclear power plants have steam generators that contain many stainless steel heat-exchanger tubes. Cracks initiate and grow in the tubes due to a stress-corrosion mechanism over time. Periodic inspections of the tubes are used to detect cracks. Consider a fleet of steam generators having a total of $n = 20{,}000$ tubes. One crack was detected after the first year of operation, which was followed by another crack during the second year and six more cracks during the third year. The data are interval-censored as the exact initiation times are unknown. A prediction interval was needed for the number of tubes that would crack from the end of the third year to the end of the tenth year.

1.3. Bearing-Cage Data

The bearing-cage failure-time data are from Abernethy et al. (1983) and are provided in the online supplementary materials. Groups of aircraft engines employing this bearing cage were put into service over time (staggered entry). At the data freeze date, 6 bearing-cage failures had occurred while the remaining 1697 units with various service times were still in service (multiple right-censored data). To assure that a sufficient number of spare parts would be available to repair the aircraft engine in a timely manner, management requested a prediction interval for the number of bearing-cages that would fail in the next year, assuming 300 hours of service for each aircraft.

The purpose of this article is to show how to construct prediction intervals for the number of future events from an on-going time-to-event process, investigate the properties of different prediction methods, and give recommendations on which methods to use.

This article is organized as follows. Section 2 provides concepts and background for prediction inference. Section 3 describes the single-cohort within-sample prediction problem. Section 4 defines how the within-sample prediction is irregular and demonstrates that the plug-in method fails to provide an

CONTACT Qinglong Tian qltian@iastate.edu Department of Statistics, Iowa State University, Ames, IA 50011.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.
© 2020 American Statistical Association

asymptotically correct prediction interval. Section 5 describes the calibration method for prediction intervals and establishes its asymptotic correctness. Section 6 presents two other prediction interval methods based on predictive distributions. The first one is a general method using parametric bootstrap samples, while the second method is inspired by generalized pivotal quantities and applies to a log-location-scale family of distributions. Section 7 extends the single-cohort within-sample prediction to the multiple-cohort problem. Section 8 compares different prediction methods, through simulation, while Section 9 applies the prediction methods to the motivating examples. Section 10 discusses the choice of distribution for the time-to-event process and addresses the issue of distribution misspecification. Section 11 gives recommendations and describes potential areas for future research.

2. Background

In a general prediction problem, denote the observable data by $\boldsymbol{D}_n$ and the future random variable by $Y_n \equiv Y$; while generic for now, later this article will focus on the within-sample prediction where Y is a count. The conditional cdf for Y given $\boldsymbol{D}_n$ is denoted by $G_n(\cdot|\boldsymbol{D}_n;\boldsymbol{\theta}) \equiv G(\cdot|\boldsymbol{D}_n;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. The goal is to make inference for Y through a prediction interval, as a useful tool for quantifying uncertainty in prediction.

2.1. Prediction Intervals

When parameters in $\boldsymbol{\theta}$ are known, the one-sided upper $100(1-\alpha/2)\%$ prediction bound $\tilde{Y}_{1-\alpha/2}$ is defined as the $100(1-\alpha/2)\%$ quantile of the conditional cdf for Y, which is

$$\begin{aligned}\tilde{Y}_{1-\alpha/2} &= \inf\{y \in \mathbb{R} : G(y|\boldsymbol{D}_n;\boldsymbol{\theta}) \\ &= \Pr(Y \le y|\boldsymbol{D}_n,\boldsymbol{\theta}) \ge 1-\alpha/2\}, \end{aligned} \quad (1)$$

and the one-sided lower $100(1-\alpha/2)\%$ prediction bound may be defined as

$$\underset{\sim}{Y}_{1-\alpha/2} = \sup\{y \in \mathbb{R} : \Pr(Y \ge y|\boldsymbol{D}_n,\boldsymbol{\theta}) \ge 1-\alpha/2\}, \quad (2)$$

where this modification of the usual $\alpha/2$ quantile of Y ensures that $\Pr(Y \ge \underset{\sim}{Y}_{1-\alpha/2}|\boldsymbol{D}_n,\boldsymbol{\theta})$ is at least $100(1-\alpha/2)\%$ when Y is a discrete random variable. We may obtain an equal-tail $100(1-\alpha)\%$ prediction interval (approximate when Y is a discrete random variable) by combining these two prediction bounds.

In most applications, equal-tail prediction intervals are preferred over unequal ones, even though it is sometimes possible to find a narrower prediction interval with unequal tail probabilities. This is because the equal-tail prediction interval can be naturally decomposed into a practical one-sided upper prediction bound and a lower prediction bound where the separate consideration of one-sided bounds is needed when the cost of being outside the prediction bound is much higher on one side than the other.

When the parameters in $\boldsymbol{\theta}$ are unknown, an estimation of $\boldsymbol{\theta}$ from the observed data $\boldsymbol{D}_n$ is required. The plug-in method, also known as the naive or estimative method (see Section 2.3), is to replace $\boldsymbol{\theta}$ with a consistent estimator $\widehat{\boldsymbol{\theta}}_n$ in the prediction bounds (1) and (2). The $100(1-\alpha)\%$ plug-in upper prediction bound is then $\tilde{Y}^{\mathrm{PL}}_{1-\alpha} = \inf\{y \in \mathbb{R} : G(y|\boldsymbol{D}_n;\widehat{\boldsymbol{\theta}}_n) \ge 1-\alpha\}$ while the $100(1-\alpha)\%$ plug-in lower prediction bound is $\underset{\sim}{Y}^{\mathrm{PL}}_{1-\alpha/2} = \sup\{y \in \mathbb{R} : \Pr(Y \ge y|\boldsymbol{D}_n,\widehat{\boldsymbol{\theta}}_n) \ge 1-\alpha\}$.

2.2. Coverage Probability

Besides the plug-in method, other methods for computing prediction bounds or intervals are available. Let $\mathrm{PI}(1-\alpha)$ generically denote a prediction interval (or bound) of a nominal coverage level $100(1-\alpha)\%$, where researchers would like the probability of Y falling within the interval to be (or close to) $1-\alpha$ (i.e., $\Pr[Y \in \mathrm{PI}(1-\alpha)] = 1-\alpha$).

To be clear, there are two possible types of coverage probability: conditional coverage probability and unconditional (overall) coverage probability. The conditional coverage probability of a particular $\mathrm{PI}(1-\alpha)$ method is defined as

$$\mathrm{CP}[\mathrm{PI}(1-\alpha)|\boldsymbol{D}_n;\boldsymbol{\theta}] = \Pr[Y \in \mathrm{PI}(1-\alpha)|\boldsymbol{D}_n;\boldsymbol{\theta}],$$

where $\Pr(\cdot|\boldsymbol{D}_n;\boldsymbol{\theta})$ denotes the conditional probability of Y given the observable data $\boldsymbol{D}_n$. The conditional coverage probability $\mathrm{CP}[\mathrm{PI}(1-\alpha)|\boldsymbol{D}_n;\boldsymbol{\theta}]$ is a random variable because it is a function of the data $\boldsymbol{D}_n$. The unconditional coverage probability of a prediction interval method can be obtained by taking an expectation with respect to the data $\boldsymbol{D}_n$ and it is defined as

$$\mathrm{CP}[\mathrm{PI}(1-\alpha);\boldsymbol{\theta}] = \mathbf{E}\left\{\Pr[Y \in \mathrm{PI}(1-\alpha)|\boldsymbol{D}_n;\boldsymbol{\theta}]\right\}.$$

The unconditional coverage probability is a fixed property of a prediction method and, as such, can be most readily studied and used to compare alternative prediction interval methods. We focus on unconditional coverage probability in this article and use the term coverage probability to refer to the unconditional probability, unless stated otherwise.

We say a prediction method is exact if $\mathrm{CP}[\mathrm{PI}(1-\alpha);\boldsymbol{\theta}] = 1-\alpha$ holds. If $\mathrm{CP}[\mathrm{PI}(1-\alpha);\boldsymbol{\theta}]$ converges to $1-\alpha$ as the sample size n increases, we say the corresponding prediction method is asymptotically correct. When Y is a discrete random variable, however, asymptotic correctness and exactness may not generally hold or be possible for a prediction interval method, due to the discreteness in the distribution of Y.

2.3. Related Literature

Extensive research exists regarding some methods for computing prediction intervals. While the plug-in method has been criticized for ignoring the uncertainty in $\widehat{\boldsymbol{\theta}}_n$, this method is often widely viewed as being asymptotically correct (related to "regular predictions" described in Section 4.1). For example, Cox (1975), Beran (1990), and Hall, Peng, and Tajvidi (1999) showed that the coverage probability of the plug-in method has an accuracy of $O(n^{-1})$ for a continuous predictand under certain conditions. In Section 4, we show, however, that the plug-in method is not asymptotically correct in the context of within-sample prediction.

Section 5 presents a calibration method for within-sample prediction intervals. Cox (1975) originally proposed the calibration idea to improve on the plug-in method and also provided analytical forms for prediction intervals based on general

asymptotic expansions. Atwood (1984) used a similar method. Beran (1990) employed bootstrap in the calibration method, avoiding the complicated analytical expressions. Escobar and Meeker (1999) described similar methods for constructing prediction intervals for failure times and the number of future failures, based on censored life data.

This article does not specifically address Bayesian prediction methods, but the classic idea of a Bayesian predictive distribution can be extended to non-Bayesian methods and two such methods are considered in Section 6. Several authors have considered similar notions of a non-Bayesian predictive distribution (e.g., Aitchison 1975; Davison 1986; Barndorff-Nielsen and Cox 1996). Lawless and Fredette (2005) demonstrated a relationship between predictive distributions and (approximate) pivotal-based prediction intervals, including the calibration method described in Beran (1990). Fonseca, Giummolè, and Vidoni (2012) further elaborated on the relationship between predictive distributions and the calibration method. Shen, Liu, and Xie (2018) proposed a general framework to construct a predictive distribution by replacing the posterior distribution in the definition of a Bayesian predictive distribution with a confidence distribution.

3. Single Cohort Within-Sample Prediction

3.1. Within-Sample Prediction and New Sample Prediction

The term "within-sample" prediction has been used to distinguish from the more widely known "new sample" prediction. In new-sample prediction, past data are used, for example, to compute a prediction interval for the lifetime of a single unit from a new and completely independent sample. For within-sample prediction, however, the sample has not changed; the future random variable that researchers wish to predict (i.e., a count) relates to the same sample that provided the original (censored) data.

3.2. Single-Cohort Within-Sample Prediction and Plug-in Method

Let $(T_1, \ldots, T_n)$ be an unordered random sample from a parametric distribution $F(t;\boldsymbol{\theta})$ having support on the positive real line and $\boldsymbol{\theta} \in \mathbb{R}^q$. Under Type I censoring at $t_c > 0$, the available data may then be expressed by $D_i = (\delta_i, T_i^{\text{obs}}), i = 1, \ldots, n$, where $\delta_i = \text{I}(T_i \le t_c)$ is a variable indicating whether T_i is observed before the censoring time t_c, so that the actual observed variables are given as $T_i^{\text{obs}} = T_i\delta_i + t_c(1-\delta_i)$. The observed number of events (uncensored units) in the sample will be denoted by $r_n = \sum_{i=1}^n \text{I}(T_i \le t_c)$. For a future time $t_w > t_c$, let $Y_n = \sum_{i=1}^n \text{I}(T_i \in (t_c, t_w])$ denote the (future) number of values from $T_1, \ldots, T_n$, that occur in the interval $(t_c, t_w]$. The conditional distribution of Y_n is then binomial$(n-r_n, p)$ given the observed data $\boldsymbol{D}_n = (D_1, \ldots, D_n)$, where p is the conditional probability that $T_i \in (t_c, t_w]$ given that $T_i > t_c$. As a function of $\boldsymbol{\theta}$, we may define p by

$$p \equiv \pi(\boldsymbol{\theta}) = \frac{F(t_w;\boldsymbol{\theta}) - F(t_c;\boldsymbol{\theta})}{1 - F(t_c;\boldsymbol{\theta})}. \tag{3}$$

The goal is to construct a prediction interval for Y_n based on the observed data $\boldsymbol{D}_n = (D_1, \ldots, D_n)$ when $\boldsymbol{\theta}$ is unknown. This is referred to as single-cohort within-sample prediction because all the units enter the system at the same time and are homogeneous; and both the data $\boldsymbol{D}_n$ and the predictand Y_n are functions of the uncensored random sample $(T_1, \ldots, T_n)$.

Let $\widehat{\boldsymbol{\theta}}_n$ denote an estimator of $\boldsymbol{\theta}$ based on $\boldsymbol{D}_n$, then a plug-in estimator $\widehat{p}_n = \pi(\widehat{\boldsymbol{\theta}}_n)$ of the conditional probability p follows from (3). Analogous to the bounds in Section 2.1, a $100(1-\alpha)\%$ plug-in lower prediction bound is defined as

$$\begin{aligned} Y_{n,1-\alpha}^{\text{PL}} &= \sup\{y \in \{0\} \cup \mathbb{Z}^+; \text{pbinom}(y-1, n-r_n, \widehat{p}_n) \le \alpha\} \\ &= \begin{cases} \text{qbinom}(\alpha, n-r_n, \widehat{p}_n), \\ \quad \text{if pbinom}(\text{qbinom}(\alpha, n-r_n, \widehat{p}_n), n-r_n, \widehat{p}_n) > \alpha, \\ \text{qbinom}(\alpha, n-r_n, \widehat{p}_n) + 1, \\ \quad \text{if pbinom}(\text{qbinom}(\alpha, n-r_n, \widehat{p}_n), n-r_n, \widehat{p}_n) = \alpha, \end{cases} \end{aligned}$$

where pbinom and qbinom are, respectively, the binomial cdf and quantile function. Similarly, the $100(1-\alpha)\%$ plug-in upper prediction bound for Y_n is defined as

$$\begin{aligned} \tilde{Y}_{n,1-\alpha}^{\text{PL}} &= \inf\{y \in \{0\} \cup \mathbb{Z}^+; \text{pbinom}(y, n-r_n, \widehat{p}_n) \ge 1-\alpha\} \\ &= \text{qbinom}(1-\alpha, n-r_n, \widehat{p}_n). \end{aligned}$$

Section 2.2 mentioned that asymptotically correct coverage may not generally be possible for prediction intervals involving a discrete predictand. However, for within-sample prediction here, prediction interval methods can be sensibly examined for properties of asymptotic correctness, which we consider in the following section. This is because discreteness in the (conditionally) binomial predictand Y_n essentially disappears in large sample sizes n, due to normal approximations.

4. The Irregularity of the Within-Sample Prediction

4.1. A Regular Prediction Problem

Under the general prediction framework described in Section 2, the conditional cdf $G_n(\cdot|\boldsymbol{D}_n;\boldsymbol{\theta})$ of a predictand Y_n given the observed data $\boldsymbol{D}_n$ is often estimated by the plug-in method as $G_n(\cdot|\boldsymbol{D}_n;\widehat{\boldsymbol{\theta}}_n)$ (also known as a predictive distribution), where $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}$ based on $\boldsymbol{D}_n$. To frame much of the literature related to the plug-in method (Section 2.3), we may define the prediction problem most often commonly related to the plug-in method as "regular" according to the following definition.

Definition 1. In the notation of Section 2, a prediction problem is called regular if

$$\sup_{y \in \mathbb{R}} |G_n(y|\boldsymbol{D}_n;\boldsymbol{\theta}) - G_n(y|\boldsymbol{D}_n;\widehat{\boldsymbol{\theta}}_n)| \xrightarrow{p} 0$$

holds as $n \to \infty$ for any consistent estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ (i.e., $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$).

Unlike coverage probability (where exactness may again not be possible for discrete predictands), the above definition reflects the underlying sense of how the plug-in method for prediction intervals is often asymptotically valid for both discrete

and continuous predictands. By the nature of many prediction problems (e.g., new sample prediction), the conditional form of cdf G_n may also not necessarily vary with n (e.g., $G_n(\cdot|\boldsymbol{D}_n;\boldsymbol{\theta}) = G(\cdot;\boldsymbol{\theta})$). Hence, in a regular prediction problem, the plug-in predictive distribution (estimated cdf) asymptotically captures the true conditional cdf of the predictand, so that differences are expected to vanish between quantiles of the true predictand Y_n and the associated plug-in prediction bounds. Further, when the predictand has a continuous and asymptotically tight conditional distribution (with probability 1), such as when the conditional cdf $G_n(\cdot|D_n;\boldsymbol{\theta}) = G(\cdot;\boldsymbol{\theta})$ of the predictand does not vary with n, then the plug-in method will be asymptotically correct.

4.2. Failure of the Plug-in Method

This section shows that the within-sample prediction problem described in Section 3 is not regular and that the plug-in method is not asymptotically valid for within-sample prediction. To avoid redundancy, the presentation of results will focus on the plug-in upper prediction bound; the lower bound is analogous by Remark 1. In the context of within-sample prediction (cf. Section 3.2), recall that the $100(1-\alpha)\%$ plug-in upper prediction bound for the future count $Y_n \equiv \sum_{i=1}^{n} \mathrm{I}(T_i \in (t_c, t_w])$ is defined as

$$\tilde{Y}^{\mathrm{PL}}_{n,1-\alpha} = \inf\{y \in \mathbb{Z}; \mathrm{pbinom}(y, n-r_n, \widehat{p}_n) \geq 1-\alpha\}.$$

The following theorem shows that the coverage probability of $\tilde{Y}^{PI}_{n,1-\alpha}$ will not correctly converge to $1-\alpha$ as n increases.

Theorem 1. Let $T_1, \ldots, T_n$ denote a random sample from a parametric distribution with cdf $F(\cdot;\boldsymbol{\theta}_0)$ (at the true value of $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \in \mathbb{R}^q$), which is observed under Type I censoring at $t_c > 0$. Suppose also that $F(t_c;\boldsymbol{\theta}_0) < 1$, $p_0 = \pi(\boldsymbol{\theta}_0) \in (0,1)$ in (3), $F(t_c;\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$, and that the conditional probability (parametric function) $p \equiv \pi(\boldsymbol{\theta})$ is continuously differentiable in a neighborhood of $\boldsymbol{\theta}_0$ with nonzero gradient $\nabla_0 \equiv \partial\pi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$. Based on the censored sample, suppose $\widehat{\boldsymbol{\theta}}_n$ is an estimator of $\boldsymbol{\theta}$ satisfying $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{MVN}(\mathbf{0}, \boldsymbol{V}_0)$, as $n \to \infty$, for a multivariate normal distribution with mean vector $\mathbf{0}$ and positive definite variance matrix $\boldsymbol{V}_0$. Then,

1. The within-sample prediction of $Y_n = \sum_{i=1}^{n} \mathrm{I}(t_c < T_i \leq t_w)$ fails to be a regular prediction problem: denoting $G_n(y|\boldsymbol{D}_n, \boldsymbol{\theta}_0) = \mathrm{pbinom}(y, n-r_n, p_0)$ as the conditional cdf of Y_n and $G_n(y|\boldsymbol{D}_n, \widehat{\boldsymbol{\theta}}_n) = \mathrm{pbinom}(y, n-r_n, \widehat{p}_n)$ as its plug-in estimator, then

$$\sup_{y\in\mathbb{R}} \left| G_n(y|\boldsymbol{D}_n, \boldsymbol{\theta}_0) - G_n(y|\boldsymbol{D}_n, \widehat{\boldsymbol{\theta}}_n) \right| \xrightarrow{d} 1 - 2\Phi_{\mathrm{nor}}(\sqrt{\nu_1}|Z_1|/2),$$

where Z_1 is a standard normal variable with cdf $\Phi_{\mathrm{nor}}(z) = \int_{-\infty}^{z} 1/\sqrt{2\pi} e^{-u^2/2} du$, $z \in \mathbb{R}$, and

$$\nu_1 \equiv \frac{[1 - F(t_c;\boldsymbol{\theta}_0)]}{p_0(1-p_0)} \nabla_0^t \boldsymbol{V}_0 \nabla_0 \in (0, \infty).$$

2. The plug-in upper prediction bound $\tilde{Y}^{\mathrm{PL}}_{n,1-\alpha}$ generally fails to have asymptotically correct coverage:

$$\lim_{n\to\infty} \Pr(Y_n \leq \tilde{Y}^{\mathrm{PL}}_{n,1-\alpha}) = \Lambda_{1-\alpha}(\nu_1) \in (0,1) \quad \text{such that}$$

$$\mathrm{sgn}\left[\Lambda_{1-\alpha}(\nu_1) - (1-\alpha)\right] = \begin{cases} 1 & \text{if } \alpha \in (1/2, 1), \\ 0 & \text{if } \alpha = 1/2, \\ -1 & \text{if } \alpha \in (0, 1/2), \end{cases}$$

where $\mathrm{sgn}(\cdot)$ is the sign function and $\Lambda_{1-\alpha}(\nu_1) \equiv \int_{-\infty}^{\infty} \Phi_{\mathrm{nor}}\left[\Phi^{-1}_{\mathrm{nor}}(1-\alpha) + z\sqrt{\nu_1}\right] d\Phi_{\mathrm{nor}}(z)$. Furthermore, $\Lambda_{1-\alpha}(\nu_1) \in [1/2, 1-\alpha)$ is a decreasing function of $\nu_1 > 0$ for a given $\alpha \in (0, 1/2)$, while $\Lambda_{1-\alpha}(\nu_1) \in (1-\alpha, 1/2]$ is increasing in $\nu_1 > 0$ for $\alpha \in (1/2, 1)$, and $\lim_{\nu_1\to\infty} \Lambda_{1-\alpha}(\nu_1) = 1/2$ holds for any $\alpha \in (0,1)$.

Remark 1. The lower plug-in bound $\tilde{Y}^{\mathrm{PL}}_{n,1-\alpha}$ behaves similarly with $\lim_{n\to\infty} \Pr(Y_n \geq \tilde{Y}^{\mathrm{PL}}_{n,1-\alpha}) = \lim_{n\to\infty} \Pr(Y_n \leq \tilde{Y}^{\mathrm{PL}}_{n,1-\alpha})$ in Theorem 1.

The proof of Theorem 1 is in the online supplementary materials. This counter-intuitive result reveals that the plug-in method should not be used to construct prediction intervals in the within-sample prediction problem, even if the sample size is large. The first part of Theorem 1 entails that plug-in estimation fails to capture the distribution of the predictand Y_n here, to the extent that the supremum difference between estimated and true distributions has a *random* limit, rather than converging to zero as in a regular prediction (see Definition 1). As a consequence, the limiting coverage probability of the plug-in bound turns out to be "off" by an amount determined by a magnitude of $\nu_1 > 0$ in Theorem 1 (part 2). For increasing values of ν_1, the coverage probability approaches 0.5, regardless of the nominal coverage level intended. An intuitive explanation for the failure of plug-in method is that, although $\widehat{p}_n$ converges consistently to p, the growing number of Bernoulli trials $n - r_n$ in Y_n offsets the improvements that larger samples may offer in estimation by $\widehat{p}_n$. In other words, when standardizing the true $1-\alpha$ quantile, say $Y_{n,1-\alpha}$, of the (conditionally binomial) predictand Y_n, one obtains a standard normal quantile $(Y_{n,1-\alpha} - p)/\sqrt{n-r_n} \approx \Phi^{-1}_{\mathrm{nor}}(1-\alpha)$ by normal approximation; however, the same standardization applied to the plug-in bound $\tilde{Y}^{\mathrm{PL}}_{n,1-\alpha}$ gives $(\tilde{Y}^{\mathrm{PL}}_{n,1-\alpha} - p)/\sqrt{n-r_n} \approx \Phi^{-1}_{\mathrm{nor}}(1-\alpha) + \sqrt{n-r_n}(\widehat{p}_n - p)$, which differs by a substantial and random amount $\sqrt{n-r_n}(\widehat{p}_n - p)$ (having a normal limit itself). Hence, validity of the plug-in method for within-sample prediction would require an estimator $\widehat{p}_n$ such that $\widehat{p}_n = p + o_p(n^{-1/2})$, which demands more than what is available from standard $\sqrt{n}$-consistency.

5. Prediction Intervals Based on Calibration

5.1. Calibrating Plug-in Prediction Bounds

Cox (1975) suggested an approximation for improving the plug-in method, which will be described next. Considering the general prediction problem (see Section 2.1), suppose a future random variable $Y \equiv Y_n$ has a conditional cdf $G_n(\cdot|\boldsymbol{D}_n;\boldsymbol{\theta}) \equiv G(\cdot|\boldsymbol{D}_n;\boldsymbol{\theta})$ given random sample $\boldsymbol{D}_n$ and $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}$ from $\boldsymbol{D}_n$. The coverage probability of the $100(1-\alpha)\%$ plug-in upper prediction bound is denoted

by $\Pr\left[G(Y|\boldsymbol{D}_n;\widehat{\boldsymbol{\theta}}_n) \le 1-\alpha\right] = 1-\alpha'$, where α' is generally different from α due to the estimation uncertainty in $\widehat{\boldsymbol{\theta}}_n$. The basic idea of the calibration method is to find a level $\alpha^\dagger$ so that the coverage probability $\Pr\left[G(Y|\boldsymbol{D}_n;\widehat{\boldsymbol{\theta}}_n) \le 1-\alpha^\dagger\right]$ is equal to (or closer to) $1-\alpha$. The resulting $100(1-\alpha^\dagger)\%$ upper plug-in prediction bound $\tilde{Y}^{\mathrm{PL}}_{n,1-\alpha^\dagger}$ is called the $100(1-\alpha)\%$ upper calibrated prediction bound. However, determination of $\alpha^\dagger$ relies on both the distribution of Y and the sampling distribution of $\widehat{\boldsymbol{\theta}}_n$, each of which depends on the unknown parameter $\boldsymbol{\theta}$. So instead, $\alpha^\dagger$ is obtained by solving the equation $\Pr_*\left[G(Y^*|\boldsymbol{D}_n;\widehat{\boldsymbol{\theta}}^*_n) \le 1-\alpha^\dagger\right] = 1-\alpha$, where $\Pr_*$ denotes bootstrap probability induced by $Y^* \sim G(\cdot|\boldsymbol{D}_n;\widehat{\boldsymbol{\theta}}_n)$ and by $\widehat{\boldsymbol{\theta}}^*_n$ as a bootstrap version of $\widehat{\boldsymbol{\theta}}_n$; for example, $\widehat{\boldsymbol{\theta}}^*_n$ may be based on a bootstrap sample $\boldsymbol{D}^*_n$ found by a parametric bootstrap applied using $\widehat{\boldsymbol{\theta}}_n$ in the role of the unknown parameter vector $\boldsymbol{\theta}$. Beran (1990) showed, that under certain conditions, instead of having a coverage error of $O(n^{-1})$, the coverage probability of the calibrated upper prediction bound improves upon the plug-in methods, for example, $\Pr\left[Y \le G^{-1}(1-\alpha^\dagger|\boldsymbol{D}_n;\widehat{\boldsymbol{\theta}}_n)\right] = 1-\alpha+O(n^{-2})$. However, such results for the validity of the calibration method cannot be applied directly to within-sample prediction because conditions in Beran (1990) entail that the prediction problem be regular (see Section 4.1), which is not true for the within-sample prediction problem (Theorem 1). Consequently, the issue of asymptotic correctness for the calibration method needs to be determined for within-sample prediction, as next considered.

5.2. The Calibration-Bootstrap Method for the Within-Sample Prediction

The general method in Beran (1990) is modified to construct a calibrated prediction interval for within-sample prediction and it is called the calibration-bootstrap method in the rest of this article. For a bootstrap sample $\boldsymbol{D}^*_n$ with r^*_n observed events (e.g., from a parametric bootstrap using $\widehat{\boldsymbol{\theta}}_n$), we define a random variable set $\left(Y^\dagger_n, n-r^*_n, \widehat{p}^*_n\right)$ where $\widehat{p}^*_n = \pi(\widehat{\boldsymbol{\theta}}^*_n)$ is the bootstrap version of $\widehat{p}_n = \pi(\widehat{\boldsymbol{\theta}}_n)$ and $Y^\dagger_n \sim \mathrm{binomial}(n-r^*_n, \widehat{p}_n)$, conditional on r^*_n.

For the $100(1-\alpha)\%$ lower prediction bound, the calibrated confidence level is

$$\alpha^\dagger_L = \sup\{u \in [0,1] : \Pr_*\left[\mathrm{pbinom}(Y^\dagger_n, n-r^*_n, \widehat{p}^*_n) \le u\right] \le \alpha\},$$

where $\Pr_*$ is the bootstrap probability induced by $\boldsymbol{D}^*_n$, and then the calibrated $100(1-\alpha)\%$ lower prediction bound is given by $\underset{\sim}{Y}^C_{n,1-\alpha} = \underset{\sim}{Y}^{\mathrm{PL}}_{n,1-\alpha^\dagger_L}$. For the $100(1-\alpha)\%$ upper prediction bound, the calibrated confidence level is

$$1-\alpha^\dagger_U = \inf\{u \in [0,1] : \Pr_*\left[\mathrm{pbinom}(Y^\dagger_n, n-r^*_n, \widehat{p}^*_n) \le u\right] \ge 1-\alpha\},$$

so that the calibrated $100(1-\alpha)\%$ upper prediction bound is $\tilde{Y}^C_{n,1-\alpha} = \tilde{Y}^{\mathrm{PL}}_{n,1-\alpha^\dagger_U}$. Here $\underset{\sim}{Y}^{\mathrm{PL}}_{n,1-\alpha}$ and $\tilde{Y}^{\mathrm{PL}}_{n,1-\alpha}$ represent lower and upper plug-in prediction bounds, respectively, as defined in Section 3.2.

The calibration-bootstrap method involves approximating the distribution of $U = \mathrm{pbinom}(Y_n, n-r_n, \widehat{p}_n)$ with the bootstrap distribution of $U^* = \mathrm{pbinom}(Y^\dagger_n, n-r^*_n, \widehat{p}^*_n)$. The bootstrap distribution of U^* is used to calibrate the plug-in method. The procedure of using the calibration-bootstrap method to construct a prediction interval is described below:

1. Compute the maximum likelihood (ML) estimate $\widehat{\boldsymbol{\theta}}_n$ using data $\boldsymbol{D}_n$ and the ML estimate $\widehat{p}_n = \pi(\widehat{\boldsymbol{\theta}}_n)$.
2. Generate a bootstrap sample $\boldsymbol{D}^*_n$ and the number of events is denoted by r^*_n.
3. Compute $\widehat{\boldsymbol{\theta}}^*_n$ and $\widehat{p}^*_n = \pi(\widehat{\boldsymbol{\theta}}^*_n)$ using the bootstrap sample $\boldsymbol{D}^*_n$.
4. Generate y^* from the distribution binomial$(n-r^*_n, \widehat{p}_n)$ and compute $u^* = \mathrm{pbinom}(y^*, n-r^*_n, \widehat{p}^*_n)$.
5. Repeat Steps 2–4 for B times and get B realizations of u^* as $\{u^*_1, \ldots, u^*_B\}$.
6. Find the α and $1-\alpha$ quantiles of $\{u^*_1, \ldots, u^*_B\}$, and denote these by u_α and $u_{1-\alpha}$, respectively. The $1-\alpha$ calibrated lower and upper prediction bounds are $\underset{\sim}{Y}^C_{n,1-\alpha} = \underset{\sim}{Y}^{\mathrm{PL}}_{n,1-u_\alpha}$ and $\tilde{Y}^C_{n,1-\alpha} = \tilde{Y}^{\mathrm{PL}}_{n,u_{1-\alpha}}$.

The pseudo-code for this algorithm is in the online supplementary materials.

Next, the calibration-bootstrap method is shown to be asymptotically correct. This requires a mild assumption on the bootstrap involved, namely that the parameter estimators $\widehat{\boldsymbol{\theta}}^*_n$ in the bootstrap world provide valid approximations for the sampling distribution of the original data estimators $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$, in large samples. More formally, let $\mathcal{L}^*_n \equiv \mathcal{L}^*_n(\boldsymbol{D}_n)$ denote the probability law of the bootstrap quantity $\sqrt{n}(\widehat{\boldsymbol{\theta}}^*_n - \widehat{\boldsymbol{\theta}}_n)$ (conditional on the data $\boldsymbol{D}_n$) and let $\mathcal{L}_n$ denote the probability law of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$. Let $\rho(\mathcal{L}_n, \mathcal{L}^*_n)$ denote the distance between these distributions under any metric $\rho(\cdot,\cdot)$ that metricizes the topology of weak convergence (e.g., the Prokhorov metric). Also, in the bootstrap recreation, the probability $\Pr_*(T^*_1 \le t_c)$ that a bootstrap observation T^*_1 is observed before the censoring time t_c should be a consistent estimator of $F(t_c;\boldsymbol{\theta})$ (e.g., $\Pr_*(T^*_1 \le t_c) = F(t_c;\widehat{\boldsymbol{\theta}}_n)$ would hold as a natural estimator under a parametric bootstrap).

Theorem 2. Under the conditions of Theorem 1, suppose that $\rho(\mathcal{L}^*_n, \mathcal{L}_n) \xrightarrow{p} 0$ and $\Pr_*(T^*_1 \le t_c) \xrightarrow{p} F(t_c;\boldsymbol{\theta}_0)$ as $n \to \infty$. Then, the $100(1-\alpha)\%$ calibrated upper and lower prediction bounds, respectively $\tilde{Y}^C_{n,1-\alpha}$ and $\underset{\sim}{Y}^C_{n,1-\alpha}$ have asymptotically correct coverage, that is

$$\lim_{n\to\infty} \Pr(Y_n \le \tilde{Y}^C_{n,1-\alpha}) = 1-\alpha = \lim_{n\to\infty} \Pr(Y_n \ge \underset{\sim}{Y}^C_{n,1-\alpha}).$$

The proof is in the online supplementary materials. Theorem 2 and its extension in Section 7 guarantee, for example, that the calibration prediction method employed in Escobar and Meeker (1999), Hong, Meeker, and McCalley (2009), and Hong and Meeker (2010, 2013) to construct the prediction intervals for the cumulative number of events is asymptotically correct.

6. Prediction Intervals Based on Predictive Distributions

6.1. Predictive Distributions

Under the general prediction setting in Section 2, recall that the predictive distribution under the plug-in method, given

by $G(\cdot|\boldsymbol{D}_n, \widehat{\boldsymbol{\theta}}_n)$, provides an estimator of the conditional cdf $G(\cdot|\boldsymbol{D}_n; \boldsymbol{\theta})$, of the predictand Y. Quantiles of this predictive distribution can be associated with prediction bounds for Y. Generally speaking, any method that leads to a prediction bound for Y can be translated to a predictive distribution by defining the $100(1-\alpha)\%$ upper prediction bound as the $1-\alpha$ quantile of the predictive distribution (and vice versa). In this section, the strategy is to construct predictive distributions that lead to prediction bound (or interval) methods having asymptotically correct coverage for within-sample prediction.

For this purpose, it is helpful to consider a Bayesian predictive distribution, defined by

$$G_B(y|\boldsymbol{D}_n) = \int G(y|\boldsymbol{D}_n; \boldsymbol{\theta})\gamma(\boldsymbol{\theta}|\boldsymbol{D}_n)d\boldsymbol{\theta}, \tag{4}$$

where $\gamma(\boldsymbol{\theta}|\boldsymbol{D}_n)$ is a joint posterior distribution for $\boldsymbol{\theta}$. The $1-\alpha$ quantile of the Bayesian predictive distribution provides the $100(1-\alpha)\%$ upper Bayesian prediction bound. While this article does not pursue the Bayesian method, the idea of the Bayesian predictive distribution can nevertheless be used by replacing the posterior $\gamma(\boldsymbol{\theta}|\boldsymbol{D}_n)$ in (4) with an alternative distribution over parameters to similarly define non-Bayesian predictive distributions. Harris (1989) replaced the posterior distribution in (4) with the bootstrap distribution of the parameters to construct a predictive distribution while Wang, Hannig, and Iyer (2012) replaced the posterior distribution with a fiducial distribution. Shen, Liu, and Xie (2018) proposed a framework for predictive inference by replacing the posterior distribution in (4) with a confidence distribution (CD) and provided theoretical results for this CD-based predictive distribution for the case of a scalar parameter. A CD is a probability distribution that can quantify the uncertainty of an unknown parameter, where both the bootstrap distribution in Harris (1989) and the fiducial distribution in Wang, Hannig, and Iyer (2012) can be viewed as CDs; see Xie and Singh (2013) for a review of these ideas.

To summarize, a predictive distribution can be constructed by using a data-based distribution on the parameter space to replace the posterior distribution in (4). Following this idea, we aim to use draws from a joint probability distribution for the parameters such that the resulting predictive distribution can be used to construct asymptotically correct prediction bounds and intervals for within-sample prediction. In particular, we propose two ways of constructing predictive distributions, extending the framework proposed by Shen, Liu, and Xie (2018) to the within-sample prediction case. In Section 6.2, we describe a prediction method that is based on the bootstrap distribution of the parameters and it is called the direct-bootstrap method in this article. In Section 6.3, we describe another method that works specifically with the (log)-location-scale family of distributions. This method is inspired by generalized pivotal quantities (GPQ) and involves generating bootstrap samples and it is called the GPQ-bootstrap method.

6.2. The Direct-Bootstrap Method

For within-sample prediction, recall that number Y_n of events between the censoring time t_c and a future time $t_w > t_c$, given the Type I censored data $\boldsymbol{D}_n$, is binomial$(n-r_n, p)$, where r_n is the number of events observed in $\boldsymbol{D}_n$ and p is the conditional probability in (3). The direct-bootstrap method uses the distribution of a bootstrap version $\widehat{p}_n^* = \pi(\widehat{\boldsymbol{\theta}}_n^*)$ of $\widehat{p}_n = \pi(\widehat{\boldsymbol{\theta}}_n)$, which is induced by the distribution of estimates $\widehat{\boldsymbol{\theta}}_n^*$ from a bootstrap sample $\boldsymbol{D}_n^*$, to construct a predictive distribution. Letting $\Pr_*$ denote bootstrap probability (probability induced by a bootstrap sample $\boldsymbol{D}_n^*$), the predictive distribution constructed using direct-bootstrap method is

$$\begin{aligned} G_{Y_n}^{\mathrm{DB}}(y|\boldsymbol{D}_n) &= \int \mathrm{pbinom}(y, n-r_n, \widehat{p}_n^*)\,\Pr_*\left(d\widehat{p}_n^*\right) \\ &\approx \frac{1}{B}\sum_{b=1}^{B} \mathrm{pbinom}(y, n-r_n, \widehat{p}_b^*), \end{aligned} \tag{5}$$

where $\widehat{p}_1^*, \ldots, \widehat{p}_B^*$ are realized bootstrap versions of $\widehat{p}_n$ from B independently generated bootstrap samples $\boldsymbol{D}_n^{*(1)}, \ldots, \boldsymbol{D}_n^{*(B)}$, and B is the number of bootstrap samples. The $100(1-\alpha)\%$ lower and upper prediction bounds using the direct-bootstrap method are then

$$\begin{aligned} Y_{n,1-\alpha}^{\mathrm{DB}} &= \sup\left\{y \in \{0\} \cup \mathbb{Z}^+ : G_{Y_n}^{\mathrm{DB}}(y-1|\boldsymbol{D}_n) \le \alpha\right\}, \\ \tilde{Y}_{n,1-\alpha}^{\mathrm{DB}} &= \inf\left\{y \in \{0\} \cup \mathbb{Z}^+ : G_{Y_n}^{\mathrm{DB}}(y|\boldsymbol{D}_n) \ge 1-\alpha\right\}. \end{aligned} \tag{6}$$

6.3. The GPQ-Bootstrap Method

This section focuses on the log-location-scale distribution family and develops another method to construct a predictive distribution through approximate GPQs. Suppose $(T_1, \ldots, T_n)$ is an iid random sample from a log-location-scale distribution

$$F(t; \mu, \sigma) = \Phi\left[\frac{\log(t)-\mu}{\sigma}\right], \tag{7}$$

where $\Phi(\cdot)$ is a known cdf that is free of parameters. For example, if $\Phi(\cdot)$ is the standard normal cdf $\Phi_{\mathrm{nor}(\cdot)}$, then T_1 has the log-normal distribution.

Hannig, Iyer, and Patterson (2006) described methods for constructing GPQs and outlined the relationship between GPQs and fiducial inference. Applying these ideas, GPQs can be defined for the parameters (μ, σ) in the log-location-scale model as follows. If $\mathbb{S}$ is a complete or Type II censored independent sample from a log-location-scale distribution, a set of GPQs for (μ, σ) under $\mathbb{S}$ is given by

$$\mu_n^{**} = \widehat{\mu}_n + \left(\frac{\mu - \widehat{\mu}_n^{\mathbb{S}^*}}{\widehat{\sigma}_n^{\mathbb{S}^*}}\right)\widehat{\sigma}_n \quad \text{and} \quad \sigma_n^{**} = \left(\frac{\sigma}{\widehat{\sigma}_n^{\mathbb{S}^*}}\right)\widehat{\sigma}_n, \tag{8}$$

where $\mathbb{S}^*$ denotes an independent copy of the sample $\mathbb{S}$, and $(\widehat{\mu}_n, \widehat{\sigma}_n)$ and $(\widehat{\mu}_n^{\mathbb{S}^*}, \widehat{\sigma}_n^{\mathbb{S}^*})$ denote the ML estimators of (μ, σ) computed from $\mathbb{S}$ and $\mathbb{S}^*$, respectively. These GPQs induce a distribution over the parameter space (μ, σ) based on data estimates $(\widehat{\mu}_n, \widehat{\sigma}_n)$ and, due to the fact that $[(\mu - \widehat{\mu}_n)/\sigma, \widehat{\sigma}_n/\sigma]$ are pivotal quantities based on a complete or Type II censored sample $T_1, \ldots, T_n$ from the log-location-family, the distribution of $[(\mu - \widehat{\mu}_n^{\mathbb{S}^*})/\widehat{\sigma}_n^{\mathbb{S}^*}, \sigma/\widehat{\sigma}_n^{\mathbb{S}^*})]$ in (8) can be directly approximated by simulation.

GPQs can also, in some applications, be used to construct confidence intervals when an exact pivot is unavailable. Notice that, while the quantities in (8) are GPQs for log-location-scale family based on complete or Type II censored data, these

are no longer GPQs with Type I censored data, where exact GPQs technically fail to exist. This is because the distribution of $[(\mu-\widehat{\mu}_n)/\widehat{\sigma}_n, \sigma/\widehat{\sigma}_n]$ depends on the unknown event probability $F(t_c;\mu,\sigma)$ before the censoring time t_c under Type I censoring, which applies also to $\left[(\mu-\widehat{\mu}_n^{\mathbb{S}^*})/\widehat{\sigma}_n^{\mathbb{S}^*}, \sigma/\widehat{\sigma}_n^{\mathbb{S}^*}\right]$.

However, the formula in (8) can be used to provide a joint approximate GPQ distribution under Type I censoring. Letting $\widehat{\boldsymbol{\theta}}_n^* = (\widehat{\mu}_n^*, \widehat{\sigma}_n^*)$ denote a bootstrap version of $\widehat{\boldsymbol{\theta}}_n = (\widehat{\mu}_n, \widehat{\sigma}_n)$, (8) is extended to define a joint approximate GPQ distribution as the bootstrap distribution of $\widehat{\boldsymbol{\theta}}_n^{**} = (\widehat{\mu}_n^{**}, \widehat{\sigma}_n^{**})$, where

$$\widehat{\mu}_n^{**} = \widehat{\mu}_n + \left(\frac{\widehat{\mu}_n - \widehat{\mu}_n^*}{\widehat{\sigma}_n^*}\right)\widehat{\sigma}_n \quad \text{and} \quad \widehat{\sigma}_n^{**} = \left(\frac{\widehat{\sigma}_n}{\widehat{\sigma}_n^*}\right)\widehat{\sigma}_n. \tag{9}$$

The above definition of $\widehat{\boldsymbol{\theta}}_n^{**}$ also follows by using the bootstrap distribution of $\left[(\widehat{\mu}_n - \widehat{\mu}_n^*)/\widehat{\sigma}_n^*, \widehat{\sigma}_n/\widehat{\sigma}_n^*\right]$ to approximate the sampling distribution of $[(\mu-\widehat{\mu}_n)/\widehat{\sigma}_n, \sigma/\widehat{\sigma}_n]$ and linearly solving for (μ,σ). Then using $\widehat{\boldsymbol{\theta}}_n^{**} = (\widehat{\mu}_n^{**}, \widehat{\sigma}_n^{**})$ instead of $\widehat{\boldsymbol{\theta}}_n^* = (\widehat{\mu}_n^*, \widehat{\sigma}_n^*)$, a predictive distribution can be defined by using the same procedure that defined the predictive distribution in (5). Namely, by defining a random variable $\widehat{p}_n^{**} \equiv \pi(\widehat{\boldsymbol{\theta}}_n^{**})$ from (3) with a bootstrap distribution induced by $\widehat{\boldsymbol{\theta}}_n^{**} = (\widehat{\mu}_n^{**}, \widehat{\sigma}_n^{**})$, the predictive distribution for Y_n using the GPQ-bootstrap method is given by

$$\begin{aligned} G_{Y_n}^{\mathrm{GPQ}}(y|\boldsymbol{D}_n) &= \int \mathrm{pbinom}(y, n-r_n, \widehat{p}_n^{**})\,\mathrm{Pr}_*\left(d\widehat{p}_n^{**}\right) \\ &\approx \frac{1}{B}\sum_{b=1}^{B} \mathrm{pbinom}(y, n-r_n, \widehat{p}_b^{**}), \end{aligned}$$

where $\widehat{p}_1^{**}, \ldots, \widehat{p}_B^{**}$ are computed from realized bootstrap samples. The $100(1-\alpha)\%$ lower and upper prediction bounds using GPQ-bootstrap method can be obtained by replacing the predictive distribution $G_{Y_n}^{\mathrm{DB}}(\cdot|\cdot)$ with $G_{Y_n}^{\mathrm{GPQ}}(\cdot|\cdot)$ in (6).

6.4. Coverage Probability of the Proposed Methods

This section shows that both the direct-bootstrap method (Section 6.2) and the GPQ-bootstrap method (Section 6.3) produce asymptotically correct prediction bounds/intervals for the future count Y_n. Hence, these two methods yield asymptotically valid inference for within-sample prediction of Y_n, as does the calibration-bootstrap method (Theorem 2, Section 5), but not by the standard plug-in method (Theorem 1, Section 4).

Theorem 3. Under the same conditions as Theorem 2,

1. The $100(1-\alpha)\%$ upper and lower prediction bounds using the direct-bootstrap method, respectively, $\tilde{Y}_{n,1-\alpha}^{\mathrm{DB}}$ and $\underset{\sim}{Y}_{n,1-\alpha}^{\mathrm{DB}}$, have asymptotically correct coverage. That is,
$$\lim_{n\to\infty}\Pr(Y_n \le \tilde{Y}_{n,1-\alpha}^{\mathrm{DB}}) = 1-\alpha = \lim_{n\to\infty}\Pr(Y_n \ge \underset{\sim}{Y}_{n,1-\alpha}^{\mathrm{DB}}).$$
2. If the parametric distribution $F(\cdot;\mu,\sigma)$ belongs to the log-location-scale distribution family (7), with standard cdf $\Phi(\cdot)$ differentiable on $\mathbb{R}$, the $100(1-\alpha)\%$ upper and lower prediction bounds using the GPQ-bootstrap method, respectively, $\tilde{Y}_{n,1-\alpha}^{\mathrm{GPQ}}$ and $\underset{\sim}{Y}_{n,1-\alpha}^{\mathrm{GPQ}}$, have asymptotically correct coverage. That is,
$$\lim_{n\to\infty}\Pr(Y_n \le \tilde{Y}_{n,1-\alpha}^{\mathrm{GPQ}}) = 1-\alpha = \lim_{n\to\infty}\Pr(Y_n \ge \underset{\sim}{Y}_{n,1-\alpha}^{\mathrm{GPQ}}).$$

The proof of Theorem 3 is in the online supplementary materials.

7. Multiple Cohort Within-Sample Prediction

7.1. Multiple Cohort Data

So far, the focus has been on the within-sample prediction for single-cohort data. Multiple cohort data, however, are more common in applications. In this section, the results from single-cohort data are extended to multiple-cohort data.

In multiple-cohort data (e.g., the bearing cage data of Section 1), units from different cohorts are placed into service at different times. The multiple-cohort data $\mathbb{D}$ can be seen as a collection of several single-cohort datasets as $\mathbb{D} = \{\boldsymbol{D}_{n_s}, s = 1,\ldots,S\}$, where S is the number of cohorts and n_s is the number of units in the cohort s (sometimes, with no grouping, many cohorts have size 1). Within each cohort $\boldsymbol{D}_{n_s} = (D_{s,1},\ldots,D_{s,n_s})$, we may express an observation involved as $D_{s,i} = (\delta_i^s, T_i^{\mathrm{obs},s})$, for $T_i^{\mathrm{obs},s} = T_i^s\delta_i^s + (1-\delta_i^s)t_c^s$, where T_i^s is a random variable from a parametric distribution $F(\cdot;\boldsymbol{\theta})$, t_c^s is the censoring time for cohort s, and $\delta_i^s = \mathrm{I}(T_i^s \le t_c^s)$ is a random variable indicating whether a unit's value (e.g., failure time) is less than the censoring time t_c^s. Given the multiple-cohort data $\mathbb{D}$, the number of observed events (e.g., failures) within cohort s is defined as $r_{n_s} = \sum_{i=1}^{n_s}\mathrm{I}(T_i^s \le t_c^s), s = 1,\ldots,S$, where the total number of units is $n = \sum_{s=1}^{S} n_s$. The predictand in the multiple-cohort data is the total number of events that will occur in a future time window of length Δ and it is denoted by $Y_n = \sum_{s=1}^{S}\sum_{i=1}^{n_s}\mathrm{I}(t_c^s < T_i^s \le t_w^s)$, where $t_w^s = t_c^s + \Delta$ for $s = 1,\ldots,S$.

Within each cohort $s = 1,\ldots,S$, the number $Y_s = \sum_{i=1}^{n_s}\mathrm{I}(t_c^s < T_i^s \le t_w^s)$ of future events has a binomial distribution. As in Section 3, the conditional distribution of Y_s is binomial$(n-r_{n_s}, p_s)$, where p_s is defined as

$$p_s \equiv \pi_s(\boldsymbol{\theta}) = \frac{F(t_w^s;\boldsymbol{\theta}) - F(t_c^s;\boldsymbol{\theta})}{1 - F(t_c^s;\boldsymbol{\theta})}, \quad s = 1,\ldots,S.$$

Consequently, the predictand $Y_n = \sum_{s=1}^{S} Y_s$ has a Poisson-binomial distribution with probability vector $\boldsymbol{p} = (p_1,\ldots,p_S)$ and weight vector $\boldsymbol{w} = (n_1 - r_{n_1},\ldots,n_S - r_{n_S})$. We denote this Poisson-binomial distribution by Poibin$(\boldsymbol{p},\boldsymbol{w})$, where the cdf of the Poisson-binomial distribution is denoted by ppoibin$(\cdot,\boldsymbol{p},\boldsymbol{w})$ and the quantile function is denoted by qpoibin$(\cdot,\boldsymbol{p},\boldsymbol{w})$; these functions are available in the `poibin` R package (described in Hong (2013)).

If $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}$ based on the multiple-cohort data $\mathbb{D}$, an estimator $\widehat{\boldsymbol{p}} = (\widehat{p}_n^1,\ldots,\widehat{p}_n^S)$ of conditional probabilities $\boldsymbol{p}$ follows by substitution $\widehat{p}_n^s = \pi_s(\widehat{\boldsymbol{\theta}}_n), s = 1,\ldots,S$, similar to the single-cohort case. Then, the $100(1-\alpha)\%$ plug-in lower and upper prediction bounds for Y_n are

$$\begin{aligned} \underset{\sim}{Y}_{n,1-\alpha}^{\mathrm{PL}} &= \sup\{y \in \{0\}\cup\mathbb{Z}^+ : \mathrm{ppoibin}\left(y-1,\widehat{\boldsymbol{p}},\boldsymbol{w}\right) \le \alpha\}, \\ &= \begin{cases} \mathrm{qpoibin}(\alpha,\widehat{\boldsymbol{p}},\boldsymbol{w}), \\ \quad \text{if } \mathrm{pbinom}(\mathrm{qpoibin}(\alpha,\widehat{\boldsymbol{p}},\boldsymbol{w}),\widehat{\boldsymbol{p}},\boldsymbol{w}) > \alpha, \\ \mathrm{qpoibin}(\alpha,\widehat{\boldsymbol{p}},\boldsymbol{w}) + 1, \\ \quad \text{if } \mathrm{pbinom}(\mathrm{qpoibin}(\alpha,\widehat{\boldsymbol{p}},\boldsymbol{w}),\widehat{\boldsymbol{p}},\boldsymbol{w}) = \alpha, \end{cases} \\ \tilde{Y}_{n,1-\alpha}^{\mathrm{PL}} &= \inf\{y \in \{0\}\cup\mathbb{Z}^+ : \mathrm{ppoibin}(y,\widehat{\boldsymbol{p}},\boldsymbol{w}) \ge 1-\alpha\} \\ &= \mathrm{qpoibin}(1-\alpha,\widehat{\boldsymbol{p}},\boldsymbol{w}). \end{aligned}$$

Similar to the single-cohort case (Theorem 1), the plug-in method also fails to provide an asymptotically correct coverage probability under multiple-cohort data; see the online supplementary materials.

7.2. The Calibration-Bootstrap Method for Multiple Cohort Data

Formulating prediction bounds using the calibration-bootstrap method first requires simulation of bootstrap samples, where each bootstrap sample $\mathbb{D}^*$ matches the original data in terms of the number S of cohorts as well as their respective sizes n_s and censoring times t_c^s, $s = 1, \ldots, S$. The bootstrap version of the estimator $\widehat{\boldsymbol{p}} = (\widehat{p}_n^1, \ldots, \widehat{p}_n^S)$ is $\widehat{\boldsymbol{p}}^* = (\widehat{p}_n^{1,*}, \ldots, \widehat{p}_n^{S,*})$ from each bootstrap sample $\mathbb{D}^*$. Additionally, the number of events (e.g., failures) in the bootstrap sample, grouped by cohort, is $(r_{n_1}^*, \ldots, r_{n_S}^*)$, from which we denote a bootstrap future count by $Y_n^\dagger \sim \text{Poibin}(\widehat{\boldsymbol{p}}; \boldsymbol{w}^*)$ based on a weight vector from the bootstrap sample as $\boldsymbol{w}^* = (n_1 - r_{n_1}^*, \ldots, n_S - r_{n_S}^*)$. The bootstrap variable set $(Y_n^\dagger, \widehat{\boldsymbol{p}}^*, \boldsymbol{w}^*)$ is then applied into a Poisson-binomial cdf and then leads to a transformed random variable $U^* = \text{ppoibin}(Y_n^\dagger, \widehat{\boldsymbol{p}}^*, \boldsymbol{w}^*) \in [0, 1]$ for deriving calibrated confidence levels $\alpha_L^\dagger$ and $\alpha_U^\dagger$ in the same way as in the single-cohort situation. Then, the $100(1-\alpha)\%$ calibrated lower prediction bound is $\underset{\sim}{Y}_{n,1-\alpha}^C = \underset{\sim}{Y}_{n,1-\alpha_L^\dagger}^{\text{PL}}$ and the similar upper prediction bound version is $\tilde{Y}_{n,1-\alpha}^C = \tilde{Y}_{n,1-\alpha_U^\dagger}^{\text{PL}}$.

The calibration-bootstrap method remains asymptotically correct for multiple-cohort within-sample prediction. The multiple-cohort extensions of Theorem 2 and the algorithm are in the online supplementary materials.

7.3. The Direct- and GPQ-Bootstrap Methods for Multiple Cohort Data

For multiple-cohort data, constructing prediction bounds for Y_n based on the predictive-distribution-based methods also requires bootstrap data and, in particular, the distribution of a bootstrap version $\widehat{\boldsymbol{p}}^*$ of $\widehat{\boldsymbol{p}}$ as in Section 7.2. The predictive distribution from the direct-bootstrap method is

$$G_{Y_n}^{\text{DB}}(y|\mathbb{D}) = \int \text{ppoibin}(y, \widehat{\boldsymbol{p}}^*, \boldsymbol{w}) \, \text{Pr}_*(d\widehat{\boldsymbol{p}}^*) \approx \frac{1}{B} \sum_{b=1}^{B} \text{ppoibin}(y, \widehat{\boldsymbol{p}}_b^*, \boldsymbol{w}), \tag{10}$$

where $\widehat{\boldsymbol{p}}_1^*, \ldots, \widehat{\boldsymbol{p}}_B^*$ are realized bootstrap versions of $\widehat{\boldsymbol{p}}$ across independently generated bootstrap versions of multiple-cohort data (e.g., $\mathbb{D}^*$). The $100(1-\alpha)\%$ direct-bootstrap lower and upper prediction bounds for Y_n are defined as the modified α quantile and $1-\alpha$ quantile of this predictive distribution, respectively, and given by

$$\underset{\sim}{Y}_{n,1-\alpha}^{\text{DB}} = \sup\left\{y \in \{0\} \cup \mathbb{Z}^+ : G_{Y_n}^{\text{DB}}\left(y - 1|\mathbb{D}\right) \le \alpha\right\},$$
$$\tilde{Y}_{n,1-\alpha}^{\text{DB}} = \inf\left\{y \in \{0\} \cup \mathbb{Z}^+ : G_{Y_n}^{\text{DB}}\left(y|\mathbb{D}\right) \ge 1 - \alpha\right\}.$$

If $F(\cdot; \boldsymbol{\theta}) = F(\cdot; \mu, \sigma)$ belongs to the log-location-scale family as in (7), we use $\widehat{\boldsymbol{\theta}}_n^* = (\hat{\mu}_n^*, \hat{\sigma}_n^*)$ to compute approximate GPQs $\widehat{\boldsymbol{\theta}}_n^{**} = (\hat{\mu}_n^{**}, \hat{\sigma}_n^{**})$ using (9), and compute $\widehat{\boldsymbol{p}}^{**} = (\widehat{p}_n^{1,**}, \ldots, \widehat{p}_n^{S,**})$ where $\widehat{p}_n^{s,**} = \pi_s(\widehat{\boldsymbol{\theta}}_n^{**})$. Then the GPQ-bootstrap method can be implemented to obtain prediction bounds for Y_n by replacing $\widehat{\boldsymbol{p}}^*$ with $\widehat{\boldsymbol{p}}^{**}$ in the definition of the direct-bootstrap predictive distribution (10) and analogously determining prediction bounds from the quantiles of this predictive distribution. The direct- and GPQ-bootstrap methods produce asymptotically correct prediction bounds from multiple-cohort data, and the extension of Theorem 3 is provided in the online supplementary materials.

8. A Simulation Study

The purpose of this simulation study is to illustrate agreement for finite sample sizes with the theorems established in the previous sections and to provide insights into the performance of different methods in the case of finite samples. The details and results in this section are for Type I censored single-cohort data. Let the event of interest be the failure of a unit. We simulated Type I censored data using the two-parameter Weibull distribution and compared the coverage probabilities of the prediction bounds based on the plug-in, calibration-bootstrap, direct-bootstrap, and GPQ-bootstrap methods. The Weibull cdf is

$$F(t; \eta, \beta) = 1 - \exp\left[-\left(\frac{t}{\eta}\right)^\beta\right], \quad t > 0,$$

with positive scale η and shape β parameters, and can also be parameterized as

$$F(t; \mu, \sigma) = \Phi_{\text{sev}}\left[\frac{\log(t) - \mu}{\sigma}\right], \quad t > 0,$$

where $\Phi_{\text{sev}}(x) = 1 - \exp\left[-\exp(x)\right]$ is the cdf of the standard smallest extreme value distribution with $\mu = \log(\eta)$ and $\sigma = 1/\beta$. The conditions in Theorems 1–3 can be verified for Type I censored Weibull data, so that the Weibull distribution can be used to illustrate all of the aforementioned methods for within-sample prediction (e.g., the ML estimators of the Weibull parameters $\widehat{\boldsymbol{\theta}}_n = (\widehat{\mu}_n, \widehat{\sigma}_n)$ have sampling distributions with normal limits and can be validly approximated by parametric bootstrap as described in Scholz (2001)).

8.1. Simulation Setup

The factors for the simulation experiment are (i) $p_{f1} = F(t_c; \beta, \eta)$, the probability that a unit fails before the censoring time t_c; (ii) $\text{E}(r) = np_{f1}$, the expected number of failures at the censoring time t_c, where n is the total sample size (i.e., including both the censored and the uncensored observations); (iii) $d \equiv p_{f2} - p_{f1}$, the probability that a unit fails in a future time interval $(t_c, t_w]$ where $p_{f2} = F(t_w; \beta, \eta)$; (iv) $\beta = 1/\sigma$, the Weibull shape parameter. Because $\eta = \exp(\mu)$ is a scale parameter, without loss of generality, $\eta = 1$ was used in the simulation. A simulation with all combinations of the following factors levels was conducted: (i) $p_{f1} = 0.05, 0.1, 0.2$; (ii) $\text{E}(r) = 5, 15, 25, 35, 45$; (iii) $d = 0.1, 0.2$; (iv) $\beta = 0.5, 0.8, 2, 4$.

For each combination of the these four factors, 90% and 95% upper prediction bounds and 90% and 95% lower prediction bounds were constructed.

The procedure for the simulation is as follows:

1. Simulate $N = 5000$ Type I censored samples for each of the factors-level combinations of the four factors.
2. Use ML to estimate parameters β, η in each censored sample.
3. Compute prediction bounds using the different methods for each sample.
4. Compute the conditional (i.e., binomial) coverage probability for each of the prediction bounds.
5. Determine the unconditional coverage probability for each method by averaging the $N = 5000$ conditional coverage probabilities.

Within each of the $N = 5000$ simulated Type I censored samples, $B = 5000$ bootstrap samples were generated by parametric bootstrap (i.e., as a random sample from the fitted Weibull distribution with Type I censoring at t_c) and these samples were used for the calibration-bootstrap method and the two predictive-distribution-based methods. In the simulation, we excluded those samples having fewer than 2 failures to avoid estimability problems, so that all $N = 5000$ original samples and all the $N \times B = 25{,}000{,}000$ bootstrap samples in the simulation have at least 2 failures. The probability of a data sample with fewer than 2 failures for each factor-level combination is given in Table 1.

Table 1. Probability of an excluded sample (i.e., $r = 0$ or 1 failures) for different factor-level combinations.

	E(r) = 5	E(r) = 15	E(r) = 25	E(r) = 35	E(r) = 45
$p_{f1} = 0.05$	0.037	0.000	0.000	0.000	0.000
$p_{f1} = 0.1$	0.034	0.000	0.000	0.000	0.000
$p_{f1} = 0.2$	0.027	0.000	0.000	0.000	0.000

8.2. Simulation Results

A small subset of the plots displaying the complete simulation results are given here, as the results are generally consistent across the different factor-level combinations. Figure 1 shows the coverage probabilities from plug-in, calibration-bootstrap, direct-bootstrap, and GPQ-bootstrap methods when $\beta = 2$ and $d = 0.2$. The horizontal dashed line in each subplot represents the nominal confidence level. Plots for the other factor-level combinations are given in the online supplementary materials.

Some observations from the simulation results are:

1. The plug-in method fails to have asymptotically correct coverage probability. As p_{f1} decreases, which entails less information or fewer events observed before the censoring time t_c, the coverage probability deviates more from the nominal level.
2. The direct- and GPQ-bootstrap methods are close to each other in terms of coverage probabilities except when $E(r) = 5$. The calibration-bootstrap method differs considerably from the direct- and GPQ-bootstrap methods. The calibration-bootstrap method tends to be more conservative than the other bootstrap-based methods for constructing lower prediction bounds, and also is less conservative for constructing upper prediction bounds.
3. For the lower bounds, the direct- and GPQ-bootstrap methods dominate the calibration-bootstrap method. For the upper bounds, the coverage probabilities of the former two bootstrap-based methods are slightly conservative but still close to the nominal level. The calibration-bootstrap method is better than the direct- and GPQ-bootstrap methods in just a few of these upper bounds.

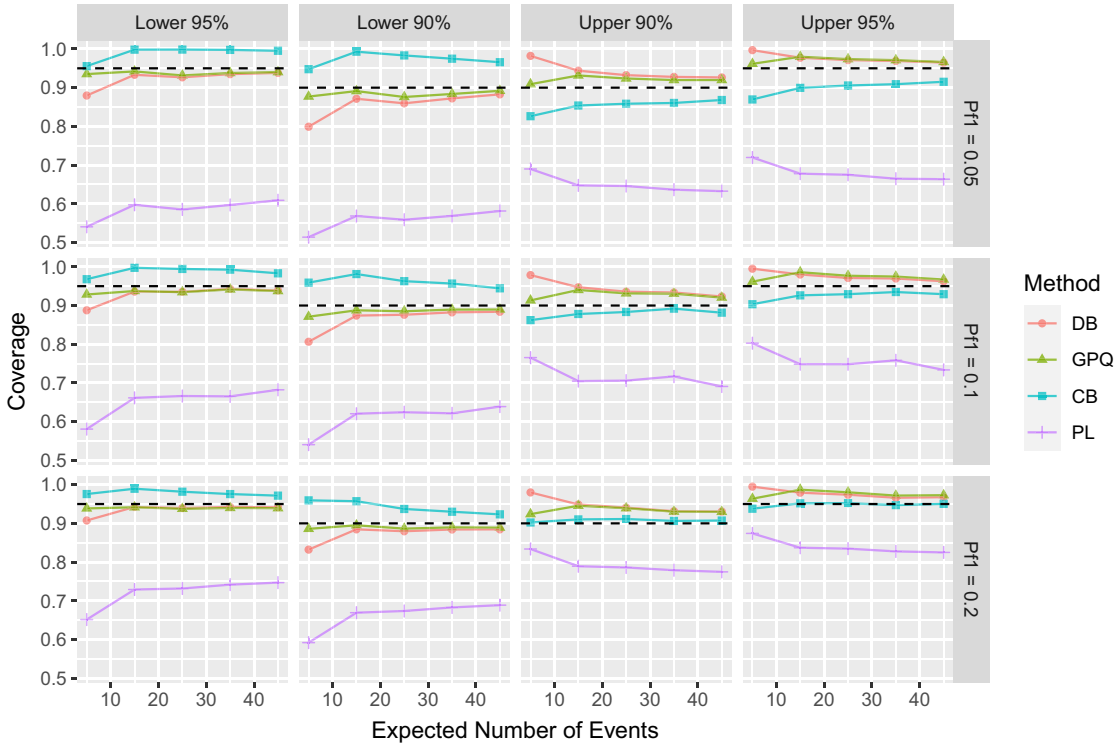

Figure 1. Coverage probabilities versus expected number of events for the direct-bootstrap (DB), GPQ-bootstrap (GPQ), calibration-bootstrap (CB), and plug-in (PL) methods when $d = p_{f2} - p_{f1} = 0.2$ and $\beta = 2$.

4. Compared with the calibration-bootstrap method, whose performance is highly related to the level of p_{f1}, the coverage probabilities of the direct- and GPQ-bootstrap methods are insensitive to the level of p_{f1}. As p_{f1} decreases, the lower prediction bound using the calibration-bootstrap method has over-coverage while the upper prediction bound has under-coverage. This implies that under heavy censoring (small p_{f1}), extremely large sample sizes n (or correspondingly large expected number of failing $\mathrm{E}(r) = np_{f1}$) are required to attain coverage probabilities close to the nominal confidence level.

From these observations, we can see that the direct- and GPQ-bootstrap methods (i.e., predictive-distribution-based methods) tend to dominate the calibration-bootstrap method in terms of the performance of the prediction bounds, even though all three methods are asymptotically valid. This is because the predictive-distribution-based methods target the one source p of parameter uncertainty in the conditional binomial$(n - r_n, p)$ distribution of the predictand Y_n (i.e., as addressed by applying bootstrap versions $\widehat{p}^*$ or $\widehat{p}^{**}$ to "smooth" estimation uncertainty for p), while the number $n - r_n$ of Bernoulli trials used in these predictive distributions matches that of the predictand. Due to its definition, however, the calibration-bootstrap method involves bootstrap approximation steps (i.e., $r_n^*, \widehat{p}^*$) for both the number r_n of failures as well as the binomial probability p. The calibration-bootstrap method essentially imposes an approximation $n - r_n^*$ for the known number $n - r_n$ of trials prescribing the predictand Y_n. As a consequence, coverages from the calibration-bootstrap method are generally less accurate than those from the predictive-distribution-based methods for within-sample prediction.

9. Application of the Methods

9.1. Examples

9.1.1. Product-A Data

The ML estimates of the Weibull shape and scale parameters are $\widehat{\beta} = 1.518$ and $\widehat{\eta} = 1152$, respectively, based on 80 failure times among 10,000 units before 48 months. Then, for the 9920 surviving units, the ML estimate of the probability that a unit will fail between 48 and 60 months of age is $\widehat{p}_n = [F(60; \widehat{\beta}, \widehat{\eta}) - F(48; \widehat{\beta}, \widehat{\eta})]/[1 - F(48; \widehat{\beta}, \widehat{\eta})] = 0.00323$. Using the ML estimates of the Weibull parameters $(\widehat{\beta}, \widehat{\eta})$, we simulate 10,000 bootstrap samples that are censored at 48 months and obtain ML estimates of (β, η) from each bootstrap sample. Based on applying these with each interval method, Table 2 gives prediction bounds for the number of failures in the next 12 months. As indicated by our results, even with a large number of failures, the plug-in method intervals can be expected to be off and are too narrow compared to the other bounds.

9.1.2. Heat Exchanger Data

In this example, there are no exact failure times in the data. That is, the data here contain limited information as there were only 8 failures among 20,000 exchanger tubes that were inspected (in censored data analysis, the informational content of data is closely related to the number of failures) and these failure times are interval-censored (not exact). The likelihood function under a Weibull model for the heat exchanger data is

$$L(\beta, \eta) = F(1; \beta, \eta)[F(2; \beta, \eta) - F(1; \beta, \eta)] \times [F(3; \beta, \eta) - F(2; \beta, \eta)]^6 [1 - F(3; \beta, \eta)]^{19992},$$

resulting in ML estimates $\widehat{\beta} = 2.531$ and $\widehat{\eta} = 66.058$. The conditional probability of a tube failing between the third and tenth year, given that tube has not failed at the end of the third year, is then estimated as $\widehat{p}_n = [F(10; \widehat{\beta}, \widehat{\eta}) - F(3; \widehat{\beta}, \widehat{\eta})]/[1 - F(3; \widehat{\beta}, \widehat{\eta})] = 0.00797$.

The ML estimates from 10,000 bootstrap samples (parametric bootstrap with censoring at 3 years) are used in the calibration-bootstrap and two predictive-distribution-based methods. However, the calibration-bootstrap method exhibits numerical instabilities with these data due to the small number of failures. To illustrate, Figure 2 shows the approximate quantile function of $U^* = \mathrm{pbinom}(Y_n^\dagger, n - r_n^*, \widehat{p}_n^*)$ used in the calibration-bootstrap method, involving the evaluation of a binomial$(n - r_n^*, \widehat{p}_n^*)$ random variable $Y_n^\dagger$ in its cdf pbinom, given the number r_n^* of failures and the estimate $\widehat{p}_n^*$ from a bootstrap sample. This quantile function is also the calibration curve, where the x-axis gives the desired confidence level $1 - \alpha$, while the y-axis gives the corresponding calibrated confidence level ($\alpha_L^\dagger$ or $1 - \alpha_U^\dagger$) to be used for determining plug-in prediction bounds (or quantiles from a binomial$(n - r_n = 19992, \widehat{p} = 0.00797)$ distribution). From Figure 2, we can see that the 0.05 and 0.1 quantiles nearly equal 0 while the 0.9 and 0.95 quantiles nearly equal 1. This creates complications in computing the prediction bounds, for example, as there is numerical instability near the 100% quantile of the binomial$(n - r_n = 19992, \widehat{p} = 0.00797)$ distribution. Consequently, 90% and 95% bounds from the calibration-bootstrap method are computationally not available (NA). Table 3 instead provides prediction bounds from the plug-in and direct- and GPQ-bootstrap methods. The plug-in prediction bounds differ substantially from the two bootstrap-based methods. Unlike the previous example (Product A data), the direct- and GPQ-bootstrap methods also differ appreciably based on the limited failure information with the heat exchanger data; we return to explore such differences in Section 9.2. The upper bounds involve a large amount of extrapolation and may not be practically meaningful other than to warn that there is a huge amount of uncertainty in the 10-year predictions.

Table 2. Product A data: prediction bounds for the number of failures in the next 12 months using different methods.

Confidence level	Bound type	Plug-in	Direct	GPQ	Calibration
95%	Lower	23	20	20	20
90%	Lower	25	23	23	23
90%	Upper	39	43	43	43
95%	Upper	42	47	47	46

9.1.3. Bearing Cage Data

In this example, staggered entry data containing multiple cohorts are considered. Table 4 gives the prediction bounds for the bearing cage dataset using 10,000 bootstrap samples. While similar in spirit to the Product-A example, the predictand here differs by having a Poisson-binomial distribution. The latter can be computed with the R package `poibin`, which is applied

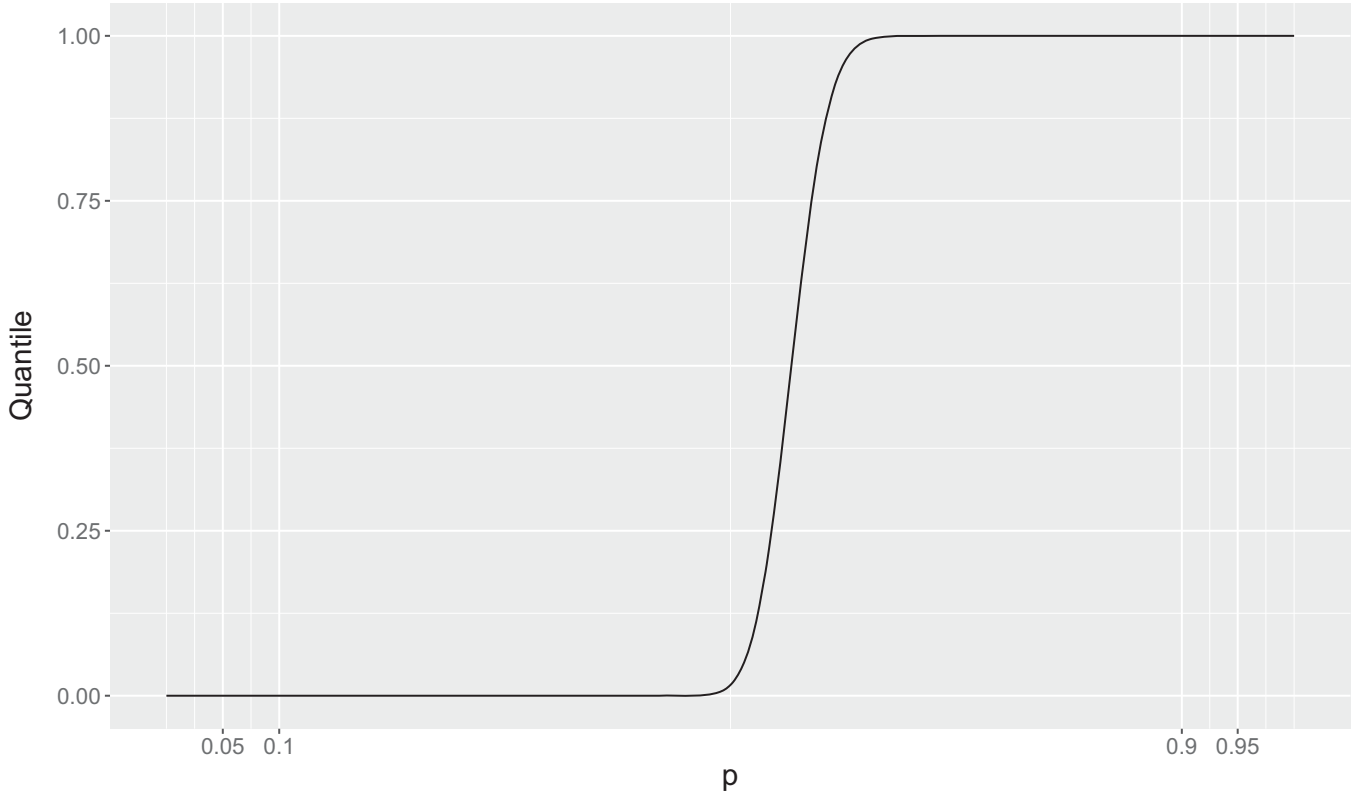

Figure 2. The quantile function of pbinom$(Y_n^\dagger, n - r_n^*, \widehat{p}_n^*)$ used for the calibration-bootstrap method with heat exchanger data.

Table 3. Heat exchanger data: prediction bounds for the number of failures in the next 7 years using different methods.

Confidence level	Bound type	Plug-in	Direct	GPQ	Calibration
95%	Lower	138	28	23	NA
90%	Lower	142	43	34	NA
90%	Upper	176	1627	888	NA
95%	Upper	180	4343	1890	NA

Table 4. Bearing cage data: prediction bounds for the number of failures in the next 300 service hours using different methods.

Confidence level	Bound type	Plug-in	Direct	GPQ	Calibration
95%	Lower	2	1	1	1
90%	Lower	2	2	2	2
90%	Upper	8	10	13	10
95%	Upper	9	12	20	12

to construct prediction bounds using methods described in Section 7.2. Table 4 gives the resulting prediction bounds for the bearing cage dataset.

9.2. Comparing the Direct- and GPQ-Bootstrap Methods

In the heat exchanger example, the prediction bounds obtained from the direct- and GPQ-bootstrap methods appear very different. This motivates us to investigate the cause of such differences in similar prediction applications involving limited information.

A general simulation setting is first described for mimicking the heat exchanger data. The heat exchanger data has two important features in that the number of events is small (i.e., 8) and so is the proportion of observed events (i.e., 0.004). Hence, in the simulation, the expected number of events $\mathrm{E}(r)$ is set to 5 while the proportion failing p_{f1} is 0.001, with a Weibull shape parameter $\beta = 2$ and scale parameter $\eta = 1$. Different levels of $d = p_{f2} - p_{f1}$ are used for the probability of events in the forecast window. The simulation results (available in the online supplementary materials) reveal that, overall, the GPQ-bootstrap method has better coverage probability than the direct-bootstrap method in this simulation setting. For the upper prediction bound, the direct-bootstrap method is generally more conservative than the GPQ-bootstrap method in terms of coverage probability, indicating that upper prediction bounds from the direct-bootstrap method are larger than the GPQ counterparts. On the other hand, the lower bound based on the direct-bootstrap method generally tends to have undercoverage compared to the GPQ-bootstrap method, suggesting also larger lower bounds from the direct-bootstrap method relative to the GPQ-bootstrap method. These patterns in the prediction bounds (i.e., with larger direct-bootstrap bounds compared to those from the GPQ-bootstrap in a setting of a limited number of events) are consistent with the prediction bounds found from the heat exchanger example. To further illustrate, Figure 3 shows the bootstrap distributions of $\widehat{p}^*$ and $\widehat{p}^{**}$ from a single Monte Carlo sample that represents the typical behavior found in this simulation setting: values of $\widehat{p}^{**}$ used in the predictive distribution of GPQ-bootstrap method tend to be smaller and more concentrated than the $\widehat{p}^*$ values used in the direct-bootstrap predictive distribution. Note that direct- and GPQ-bootstrap predictive distributions

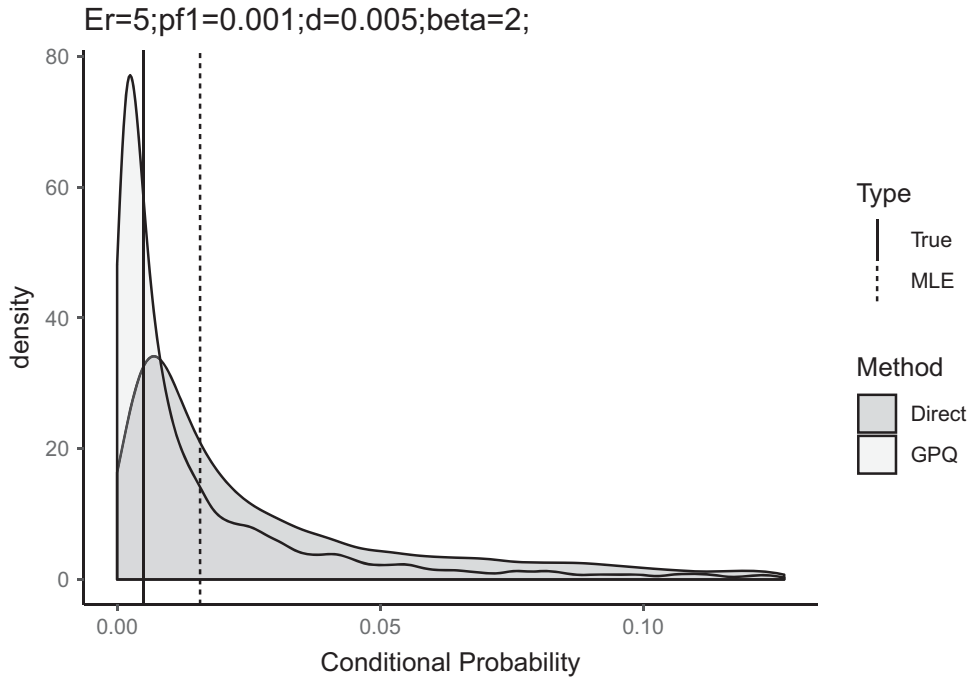

Figure 3. A representative distribution of $\widehat{p}^{*}$ and $\widehat{p}^{**}$.

are approximated by $G^{\mathrm{DB}}_{Y_n}(y|\boldsymbol{D}_n) \approx 1/B\sum_{b=1}^{B}\mathrm{pbinom}(y, n - r_n, \widehat{p}^{*}_b)$ and $G^{\mathrm{GPQ}}_{Y_n}(y|\boldsymbol{D}_n) \approx 1/B\sum_{b=1}^{B}\mathrm{pbinom}(y, n - r_n, \widehat{p}^{**}_b)$, respectively, and that direct- and GPQ-bootstrap prediction bounds correspond to quantiles from these predictive distributions. Consequently, because $\widehat{p}^{*}_b$ and $\widehat{p}^{**}_b$ are small (e.g., <0.25) while $\widehat{p}^{*}_b$ is generally larger than $\widehat{p}^{**}_b$ in Figure 3, then $G^{\mathrm{DB}}_{Y_n}(y|\boldsymbol{D}_n)$ is generally smaller than $G^{\mathrm{GPQ}}_{Y_n}(y|\boldsymbol{D}_n)$, implying quantiles from $G^{\mathrm{DB}}_{Y_n}(y|\boldsymbol{D}_n)$ can be expected to exceed those from $G^{\mathrm{GPQ}}_{Y_n}(y|\boldsymbol{D}_n)$ in data cases with a limited number of events. However, asymptotically, both $\widehat{p}^{*}_n$ and $\widehat{p}^{**}_n$ are similarly normally distributed and symmetric around $\widehat{p}_n$ (shown in online supplementary materials), so that the direct- and GPQ-bootstrap prediction bounds may be expected to behave alike in data situations with a larger number of events and larger sample sizes, as seen in Figure 1 (and in the Product A application).

10. Choice of a Distribution

Extrapolation is usually required when predicting the number of future events based on an on-going time-to-event process. For example, it may be necessary to predict the number of returns in a three-year warranty period based on field data for the first year of operation of a product. An exception arises when life can be modeled in terms of use (as opposed to time in service) and there is much variability in use rates among units in the population. The high-use units will fail early and provide good information about the upper tail of the amount-of-use return-time distribution (e.g., Hong and Meeker 2010).

When extrapolation is required, predictions can be strongly dependent on the distribution choice. In most applications, especially with heavy censoring, there is little or no useful information in the data to help choose a distribution. Then, for example, it is best to choose a failure-time distribution based on knowledge of the failure mechanism and the related physics/chemistry of failure. In important applications, this would be typically be done by consulting with experts who have such knowledge.

For example, the lognormal distribution could be justified for failure times that arise from the product of a large number of small, approximately independent positive random quantities. Examples include failure from crack initiation and growth due to cyclic stressing of metal components (e.g., in aircraft engines) and chemical degradation like corrosion (e.g., in microelectronics). These are two common applications where the lognormal distribution is often used. Gnedenko, Belyayev, and Solovyev (1969, pp. 36–37) provided mathematical justification for this physical/chemical motivation.

Based on extreme value theory, the Weibull distribution can be used to model the distribution of the minimum of a large number of approximately iid positive random variables from certain classes of distributions. For example, the Weibull distribution may provide a suitable model for the time to first failure of a large number of similar components in a system. Consider a chain with many nominally identical links and suppose that the chain is subjected cyclic stresses over time. As suggested in the previous paragraph, the number of cycles to failure for each link could be described adequately with a lognormal distribution. The chain, however, fails when the first link fails. The limiting distribution of (properly standardized) minima of iid lognormal random variables is a Type I smallest extreme value (or Gumbel) distribution. For all practical purposes, however, the Weibull distribution provides a better approximation. For further information on this result from the penultimate theory of extreme values, see Green (1976), Castillo (1988, sec. 3.11), and Gomes and de Haan (1999). Similarly, if failures are driven by the maximum of a large number of approximately iid positive

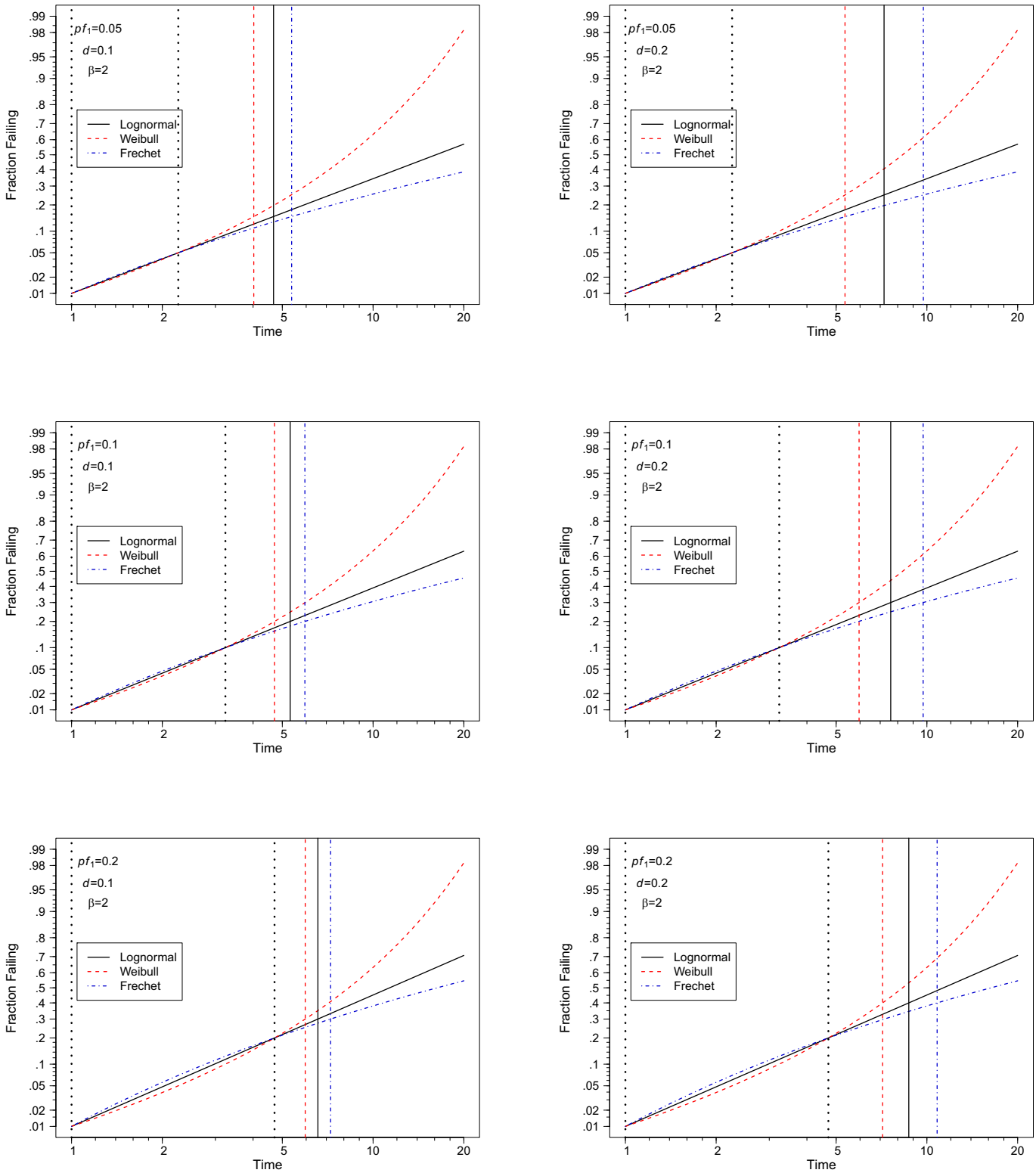

Figure 4. Distributional comparisons for $\beta = 2$. The two vertical dotted lines on the left indicate the points in time where all three distributions have the same 0.01 and p_{f1} quantiles. The three vertical lines on the right indicate the times at $p_{f2} = p_{f1} + d$ for the three distributions.

random variables, a Fréchet distribution would be suggested. The reciprocal of a Weibull random variable has a Fréchet distribution.

Of course, choosing a distribution based on failure-mechanism knowledge is not always possible. The alternative is to do sensitivity analyses, using different distributions. Figure 4 provides a comparison of the Weibull, lognormal, and Fréchet cdfs where the Weibull distribution was chosen with a shape parameter $\beta = 2$ and the other factor level combinations of p_{f1} and d used in the Section 8 simulation. The scale parameter η is determined by letting the 0.01 Weibull quantile be 1. The cdfs are plotted on lognormal probability scales where the

lognormal cdf is a straight line. The particular parameters for the lognormal and Fréchet distributions were chosen such that the distributions cross at the 0.01 and p_{f1} quantiles, simulating the range of the data where the agreement among distributions will be good. Similar plots for $\beta = 1$ and $\beta = 4$ are provided in the online supplementary materials. The Weibull distribution is always more pessimistic (conservative) than the lognormal and the Fréchet is always more optimistic than the lognormal. For example, if the true distribution is Weibull but lognormal distribution is used to fit the data, the prediction intervals, regardless of the method, will underpredict the number of events. When in doubt, the Weibull distribution is often used because it is the conservative choice.

11. Concluding Remarks

This article studies the problem of predicting the future number of events based on censored time-to-event data (e.g., failure times). This type of prediction is known as within-sample prediction. A regular prediction problem is defined for which standard plug-in estimation commonly applies, and it is shown that the within-sample prediction is not regular and that the plug-in method fails to produce asymptotically valid prediction bounds. The irregularity of within-sample prediction and the failure of the plug-in method motivated the study of the calibration method as an alternative approach for prediction bounds, though the previously established theory for calibration bounds does not apply to within-sample prediction. The calibration method is implemented via bootstrap and called calibration-bootstrap method, which is proved to be asymptotically correct (i.e., producing prediction bounds with asymptotically correct coverage). Then, turning to formulations of a predictive distribution, we study and validate two other methods to obtain prediction bounds, namely the direct-bootstrap and GPQ-bootstrap methods. All prediction methods considered can be applied to both single-cohort and multiple-cohort data.

While theoretical results show that the calibration-bootstrap method and the two predictive-distribution-based methods are all asymptotically correct, the simulation study shows that the direct-bootstrap and GPQ-bootstrap methods outperform the calibration-bootstrap method in terms of coverage probability accuracy relative to a nominal coverage level. The two predictive-distribution-based methods are also easier to implement compared to the calibration-bootstrap method, and can also be computationally more stable (e.g., heat exchanger data example). Thus, we recommend predictive distribution methods, especially the direct-bootstrap method for general applications involving within-sample prediction.

In this article, all of the units in the population were assumed to have the same time-to-event distributions. In many applications, however, units are exposed to different operating or environmental conditions, resulting in different time-to-event distributions. For example, during 1996–2000, the Firestone tires installed on Ford Explorer SUVs experienced unusually high rates of failure, where problems first arose in Saudi Arabia, Qatar, and Kuwait because of the high temperatures in those countries (see National Highway Traffic Safety Administration 2001). Having prediction intervals that use covariate information (like temperature and moisture) could be useful for manufacturers and regulators in making decisions about a possible product recall, for example. Similarly, there can be seasonality effects in time-to-event processes and within-sample predictions.

The methods described in this article can be extended to handle either constant covariates or time-varying covariates. Using calibration-bootstrap methods, Hong, Meeker, and McCalley (2009) used constant covariates to predict power-transformer failures. Despite the complicated nature of their data (random right censoring and truncation and combinations of categorical covariates with small counts in some cells), Hong, Meeker, and McCalley (2009) were able to use the fractional random-weight method (e.g., Xu et al. 2020) to generate bootstrap estimates. Shan, Hong, and Meeker (2020) used time-varying covariates to account for seasonality in two different warranty prediction applications. As mentioned by one of the referees, if there is seasonality and data from only part of one year is available, there is a difficulty. In such cases, it would be necessary to use past data on a similar process to provide information about the seasonality.

Covariate information in reliability field data has not been common, but that is changing, due to a reduction in costs and advances in sensor, communications, and storage technology. In the future, much more covariate information on various system operating/environmental variables will be available to make better predictions, as described in Meeker and Hong (2014).

Supplementary Materials

The online supplementary material provides proofs, extra simulation results, algorithms, and data.

Acknowledgments

We would like to thanks Luis A. Escobar for helpful comments on this article. We are also grateful to the editorial staff, including two reviewers, for helpful comments that improved the article.

Funding

Research was partially supported by NSF DMS-2015390.

References

Abernethy, R., Breneman, J., Medlin, C., and Reinman, G. (1983), *Weibull Analysis Handbook*, Dayton, OH: Wright-Patterson AFB. [1]

Aitchison, J. (1975), "Goodness of Prediction Fit," *Biometrika*, 62, 547–554. [3]

Atwood, C. L. (1984), "Approximate Tolerance Intervals, Based on Maximum Likelihood Estimates," *Journal of the American Statistical Association*, 79, 459–465. [3]

Barndorff-Nielsen, O. E., and Cox, D. R. (1996), "Prediction and Asymptotics," *Bernoulli*, 2, 319–340. [3]

Beran, R. (1990), "Calibrating Prediction Regions," *Journal of the American Statistical Association*, 85, 715–723. [2,3,5]

Castillo, E. (1988), *Extreme Value Theory in Engineering*, Statistical Modeling and Decision Science, New York: Academic Press. [12]

Cox, D. R. (1975), "Prediction Intervals and Empirical Bayes Confidence Intervals," *Journal of Applied Probability*, 12, 47–55. [2,4]

Davison, A. C. (1986), "Approximate Predictive Likelihood," *Biometrika*, 73, 323–332. [3]

Escobar, L. A., and Meeker, W. Q. (1999), "Statistical Prediction Based on Censored Life Data," *Technometrics*, 41, 113–124. [1,3,5]

Fonseca, G., Giummolè, F., and Vidoni, P. (2012), "Calibrating Predictive Distributions," *Journal of Statistical Computation and Simulation*, 84, 373–383. [3]

Gnedenko, B. V., Belyayev, Y. K., and Solovyev, A. D. (1969), *Mathematical Methods of Reliability Theory*, New York: Academic Press. [12]

Gomes, M. I., and de Haan, L. (1999), "Approximation by Penultimate Extreme Value Distributions," *Extremes*, 2, 71–85. [12]

Green, R. F. (1976), "Partial Attraction of Maxima," *Journal of Applied Probability*, 13, 159–163. [12]

Hall, P., Peng, L., and Tajvidi, N. (1999), "On Prediction Intervals Based on Predictive Likelihood or Bootstrap Methods," *Biometrika*, 86, 871–880. [2]

Hannig, J., Iyer, H., and Patterson, P. (2006), "Fiducial Generalized Confidence Intervals," *Journal of the American Statistical Association*, 101, 254–269. [6]

Harris, I. R. (1989), "Predictive Fit for Natural Exponential Families," *Biometrika*, 76, 675–684. [6]

Hong, Y. (2013), "On Computing the Distribution Function for the Poisson-Binomial Distribution," *Computational Statistics and Data Analysis*, 59, 41–51. [7]

Hong, Y., and Meeker, W. Q. (2010), "Field-Failure and Warranty Prediction Based on Auxiliary Use-Rate Information," *Technometrics*, 52, 148–159. [5,12]

——— (2013), "Field-Failure Predictions Based on Failure-Time Data With Dynamic Covariate Information," *Technometrics*, 55, 135–149. [5]

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," *The Annals of Applied Statistics*, 3, 857–879. [5,14]

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

Meeker, W. Q., and Hong, Y. (2014), "Reliability Meets Big Data: Opportunities and Challenges," *Quality Engineering*, 26, 102–116. [14]

National Highway Traffic Safety Administration (2001), *Engineering Analysis Report and Initial Decision Regarding EA00-023: Firestone Wilderness AT Tires*, Washington, DC: US Department of Transportation. [14]

Nelson, W. (2000), "Weibull Prediction of a Future Number of Failures," *Quality and Reliability Engineering International*, 16, 23–26. [1]

Scholz, F. (2001), "Maximum Likelihood Estimation for Type I Censored Weibull Data Including Covariates," in *ISSTECH-96-022*, Boeing Information & Support Services. [8]

Shan, Q., Hong, Y., and Meeker, W. Q. (2020), "Seasonal Warranty Prediction Based on Recurrent Event Data," *The Annals of Applied Statistics*, 14, 929–955. [14]

Shen, J., Liu, R. Y., and Xie, M.-G. (2018), "Prediction With Confidence—A General Framework for Predictive Inference," *Journal of Statistical Planning and Inference*, 195, 126–140. [3,6]

Wang, C., Hannig, J., and Iyer, H. K. (2012), "Fiducial Prediction Intervals," *Journal of Statistical Planning and Inference*, 142, 1980–1990. [6]

Xie, M.-G., and Singh, K. (2013), "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review," *International Statistical Review*, 81, 3–39. [6]

Xu, L., Gotwalt, C., Hong, Y., King, C. B., and Meeker, W. Q. (2020), "Applications of the Fractional-Random-Weight Bootstrap," *The American Statistician*, 74, 345–358. [14]

Predicting the Number of Future Events: Supplementary Material

Qinglong Tian, Fanqi Meng, Daniel J. Nordman, William Q. Meeker

Department of Statistics, Iowa State University
Ames, IA 50011

November 6, 2020

Section A outlines algorithms for computing prediction bounds with the calibration-bootstrap method in the context of single-cohort and multiple-cohort within-sample prediction. Section B provides additional simulation results for Section 9.2 of the main paper, regarding a comparison of prediction bounds from direct/GPQ-bootstrap methods with limited failure time/event information. Proofs of the main results about the asymptotic coverage properties of prediction bound methods (Theorems 1-3 from the paper) are given in Section C; these concern the single-cohort case of within-sample prediction. Extensions of these proofs to handle the multiple-cohort case are discussed in Section D. The rest of the simulation results and the Bearing Cage Data are given in Section E. Section F provides some illustrative comparisons to show how the probabilities of future events in prediction may vary by distributional model.

Section A Algorithms

Algorithm 1 describes the implementation of the calibration method with bootstrap Monte Carlo simulation. The procedure described in the main paper requires an extra layer of simulation (i.e., simulating y^* from binom$(n - r_n^*, \widehat{p}_n)$). The algorithm described below avoids this extra layer of simulation thus reducing the Monte Carlo error. But as a trade-off, more memory space is needed.

Algorithm 1: Using Bootstrap Samples to Obtain the Calibrated Upper Prediction Bound

Input: The Type I censored single cohort data:$\boldsymbol{D}_n$; The bootstrap sample size: B; The nominal level:$1 - \alpha$.

Output: The $100(1 - \alpha)\%$ upper calibrated prediction bound:$\tilde{Y}^C_{n,1-\alpha}$.

compute estimators $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(\boldsymbol{D}_n)$ and $\widehat{p}_n = \pi(\widehat{\boldsymbol{\theta}}_n)$ (e.g., by maximum likelihood);

$b \leftarrow 1$;

values_vector $\leftarrow$ NULL; prob_vector $\leftarrow$ NULL;

while $b \leq B$ **do**

  simulate the bth bootstrap sample $\boldsymbol{D}_n^{*(b)}$; the number of failures in $\boldsymbol{D}_n^{*(b)}$ is r_b^*;

  compute $\widehat{\boldsymbol{\theta}}_b^*$, as the estimate of $\boldsymbol{\theta}$ from the bootstrap sample $\boldsymbol{D}_n^{*(b)}$;

  $\widehat{p}_b^* = \pi(\widehat{\boldsymbol{\theta}}_b^*)$, where $\pi(\cdot)$ is defined in (3);

  values_vector $\leftarrow \mathbf{c}(\text{values\_vector}, \mathbf{pbinom}(0 : (n - r_b^*), n - r_b^*, \widehat{p}_b^*))$;

  prob_vector $\leftarrow \mathbf{c}(\text{prob\_vector}, \mathbf{dbinom}(0 : (n - r_b^*), n - r_b^*, \widehat{p}_n))$;

  $b \leftarrow b + 1$;

end

prob_vector $\leftarrow$ prob_vector/B;

prob_vector $\leftarrow$ prob_vector[$\mathbf{order}$(values_vector)];

empirical_cdf_y $\leftarrow \mathbf{cumsum}$(prob_vector);

empirical_cdf_x $\leftarrow \mathbf{sort}$(values_vector);

p_calibrated $\leftarrow$ empirical_cdf_x[$\mathbf{which}$(empirical_cdf_y $\geq 1 - \alpha$)[1]];

$\tilde{Y}^C_{n,1-\alpha} \leftarrow \mathbf{qbinom}(\text{p\_calibrated}, n - r, \widehat{p}_n)$;

For multiple-cohort data, the only difference is that the binomial distribution is replaced a the Poisson-binomial distribution. Algorithm 2 provides an extension of Algorithm 1 for multiple-cohort data. The functions in bold correspond to the functions available in **R**. Again only results for the upper prediction bounds are given because results for the lower prediction

bounds are similar.

Algorithm 2: Extending **Algorithm** 1 to multiple-cohort Data

Input: The Type I censoring multiple-cohort data:$\mathbb{D}$; The bootstrap sample size: B; The nominal level:$1-\alpha$.

Output: The $100(1-\alpha)\%$ upper calibrated prediction bound:$\tilde{Y}^{C}_{n,1-\alpha}$.

compute the ML estimates $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(\mathbb{D})$ and $\widehat{\boldsymbol{p}}_n = (\pi_1(\widehat{\boldsymbol{\theta}}_n), ..., \pi_S(\widehat{\boldsymbol{\theta}}_n))$.

the numbers of remaining units are $\boldsymbol{w} = (n_1 - r_{n_1}, \ldots, n_S - r_{n_S})$;

$b \leftarrow 1$;

values_vector $\leftarrow$ NULL; prob_vector $\leftarrow$ NULL;

while $b \leq B$ **do**

  simulate the bth bootstrap sample $\mathbb{D}^*_b$;

  the number of remaining units for each cohort in $\mathbb{D}^*_b$ is $\boldsymbol{w}^*_b = (n_1 - r^{*,b}_{n_1}, ..., n_S - r^{*,b}_{n_S})$;

  The total number of units at risk in the bootstrap sample $\mathbb{D}^*_b$ is $R_b = \sum_{s=1}^{S}(n_s - r^{*,b}_{n_s})$;

  compute $\widehat{\boldsymbol{\theta}}^*_b = \widehat{\boldsymbol{\theta}}^*_b(\mathbb{D}^*_b)$, the ML estimates of $\boldsymbol{\theta}$ from the bootstrap sample $\mathbb{D}^*_b$;

  compute $\widehat{\boldsymbol{p}}^*_b = (\pi_1(\widehat{\boldsymbol{\theta}}^*_b), ..., \pi_S(\widehat{\boldsymbol{\theta}}^*_b))$;

  values_vector $\leftarrow \mathbf{c}(\text{values\_vector}, \mathbf{ppoibin}(0 : R_b, \widehat{\boldsymbol{p}}^*_b, \boldsymbol{w}^*_b))$;

  prob_vector $\leftarrow \mathbf{c}(\text{prob\_vector}, \mathbf{dpoibin}(0 : R_b, \widehat{\boldsymbol{p}}_n, \boldsymbol{w}^*_b))$;

  $b \leftarrow b + 1$;

end

prob_vector $\leftarrow$ prob_vector/B;

prob_vector $\leftarrow$ prob_vector[$\mathbf{order}$(values_vector)];

empirical_cdf_y $\leftarrow \mathbf{cumsum}$(prob_vector);

empirical_cdf_x $\leftarrow \mathbf{sort}$(values_vector);

p_calibrated $\leftarrow$ empirical_cdf_x[$\mathbf{which}$(empirical_cdf_y $\geq 1-\alpha$)[1]];

$\tilde{Y}^{C}_{n,1-\alpha} \leftarrow \mathbf{qpoibin}(\text{p\_calibrated}, \widehat{\boldsymbol{p}}_n, \boldsymbol{w})$;

Section B Comparing the Direct-Bootstrap and the GPQ-Bootstrap Methods

The numerical results presented here provide additional supporting details for the simulation study described in Section 9.2, intended to compare direct-bootstrap and GPQ-bootstrap methods for data with limited event information (e.g., few failures). Table 5 gives the prediction

bounds from 10 Monte Carlo samples using $d = 0.005$, where d again denotes the (unconditional) probability of an event in a future window $(t_c, t_w]$ under the Weibull model. Figure 5

Sample	Lower 95%		Lower 90%		Upper 90%		Upper 95%	
	Direct	GPQ	Direct	GPQ	Direct	GPQ	Direct	GPQ
1	5	4	9	6	360	164	975	392
2	4	3	7	5	222	115	574	254
3	0	0	0	0	13	12	17	18
4	17	11	29	18	3851	1390	4993	3188
5	6	5	10	7	374	164	1052	417
6	9	7	15	10	991	396	3367	986
7	5	3	8	6	315	152	970	308
8	14	9	22	14	2498	770	4934	1905
9	57	80	108	120	4956	4986	4997	4997
10	4	2	6	4	264	116	825	243

Table 5: Prediction Bounds of A Few Monte Carlo Samples.

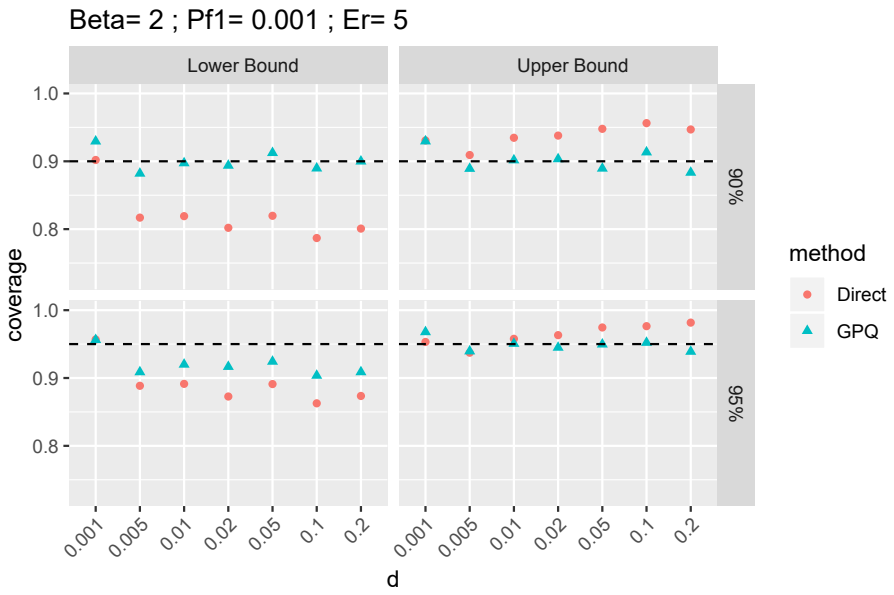

Figure 5: The Coverage Probabilities of the Direct-Bootstrap and GPQ-Bootstrap Methods.

shows the coverage probabilities of the direct-bootstrap and GPQ-bootstrap methods using different values of d. These results are referred to in the discussion of Section 9.2. Asymptotically,

both $\widehat{p}_n^*$ and $\widehat{p}_n^{**}$ are normally distributed and symmetric around $\widehat{p}_n$, as shown in Lemma 2 below.

Section C Proof of the Theorems

Here we focus on the single-cohort version of the within-sample prediction problem and the corresponding Theorems 1-3 for the prediction bound methods. Extensions of these theorems to multiple-cohort data are then discussed in Section D. The proofs of the main results require three technical lemmas (Lemmas 1-3) given next; after establishing these lemmas, proofs of Theorems 1-3 are then provided.

Recall that, under the conditions for Theorem 1, the parameter estimators $\widehat{\boldsymbol{\theta}}_n \in \mathbb{R}^q$ (e.g., ML estimators) have a normal limit, i.e., $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution to a multivariate normal $\text{MVN}(\mathbf{0}, \boldsymbol{V}_0)$. For later reference, Lemma 1 states that the bootstrap version of parameter estimators, given by $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n)$ (e.g., as generated by a parametric bootstrap), has the same normal limit under the mild consistency assumptions of Theorem 2.

Lemma 1. *Suppose conditions from Theorem 2. Letting $\mathcal{L}_n^*$ denote the bootstrap probability distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n)$ and letting $\mathcal{L}$ denote the $\text{MVN}(\mathbf{0}, \boldsymbol{V}_0)$ probability distribution, the distance $\rho(\mathcal{L}_n^*, \mathcal{L})$ between these distributions satisfies*

$$\rho(\mathcal{L}_n^*, \mathcal{L}) \xrightarrow{p} 0, \textit{as } n \to \infty,$$

under any distance $\rho(\cdot, \cdot)$ (e.g., Prokhorov distance) which metricizes weak convergence on $\mathbb{R}^q$.

Proof. If $\mathcal{L}_n$ denotes the sampling distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$, then the Theorem 1-2 conditions give $\rho(\mathcal{L}_n, \mathcal{L}) \to 0$ and $\rho(\mathcal{L}_n^*, \mathcal{L}_n) \xrightarrow{p} 0$ (by assumption). By the triangle inequality (ρ is a metric), $\rho(\mathcal{L}_n^*, \mathcal{L}) \leq \rho(\mathcal{L}_n^*, \mathcal{L}_n) + \rho(\mathcal{L}_n, \mathcal{L}) \xrightarrow{p} 0$. □

Lemma 2 next establishes certain normal limits for estimators of the conditional probability $p = \pi(\theta)$ from (3), at both the original data and bootstrap levels. Lemma 2 implies that the

bootstrap counterparts of $\sqrt{n}(\widehat{p}_n - p_0)$ converge in distribution to the same normal limit $N(0, v_0)$ as the estimator $\widehat{p}_n$ does.

Lemma 2. *Let $v_0 \equiv \nabla_0^t \boldsymbol{V}_0 \nabla_0 > 0$ with $\boldsymbol{V}_0$ and non-zero $\nabla_0 = \partial\pi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ from Theorem 1, where $\boldsymbol{\theta}_0$ contains the true parameters and $\pi(\cdot)$ is from (3). Let $\Phi_{\rm nor}(z/\sqrt{v_0})$, $z \in \mathbb{R}$, denote the cdf of a normal $N(0, v_0)$ (v_0 is the variance). Then under Theorem 2 conditions, as $n \to \infty$,*

(1) $\sqrt{n}(\widehat{p}_n - p_0) \xrightarrow{d} N(0, v_0)$ holds for the estimator $\widehat{p}_n = \pi(\widehat{\boldsymbol{\theta}}_n)$ of $p_0 = \pi(\boldsymbol{\theta}_0)$.

(2) In the direct-bootstrap method, for the bootstrap version $\sqrt{n}(\widehat{p}_n^* - \widehat{p}_n)$ of $\sqrt{n}(\widehat{p}_n - p_0)$, it holds that

$$\sup_{z\in\mathbb{R}} \left|\Pr_*(\sqrt{n}(\widehat{p}_n^* - \widehat{p}_n) \le z) - \Phi_{\rm nor}(z/\sqrt{v_0})\right| \xrightarrow{p} 0.$$

(3) In the GPQ-bootstrap method where $F(\cdot; \mu_0, \sigma_0)$ belongs to the log-location-scale family, for the approximate GPQ-based bootstrap version of $\sqrt{n}(\widehat{p}_n - p_0)$, it holds that

$$\sup_{z\in\mathbb{R}} \left|\Pr_*(\sqrt{n}(\widehat{p}_n^{**} - \widehat{p}_n) \le z) - \Phi_{\rm nor}(z/\sqrt{v_0})\right| \xrightarrow{p} 0.$$

Proof. Part 1 of Lemma 2 follows from the normal limit for $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ assumed in Theorem 1 along with the delta method as the parametric function $p = \pi(\boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}_0$. The positivity of v_0 follows because the matrix $\boldsymbol{V}_0$ is positive definite and the vector ∇_0 is non-zero.

To show the convergence in probability stated in Part 2 of Lemma 2, we use the characterization of convergence in probability through almost sure convergence along subsequences. Let $\{n_j\} \subset \{n\}$ be an arbitrary subsequence of indices. Since $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ under Theorem 1 and $\rho(\mathcal{L}_n^*, \mathcal{L}) \xrightarrow{p} 0$ by Lemma 1, there exists a further subsequence $\{n_k\} \subset \{n_j\}$ and both $\widehat{\boldsymbol{\theta}}_{n_k} \to \boldsymbol{\theta}_0$ and $\rho(\mathcal{L}_{n_k}^*, \mathcal{L}_{n_k}) \to 0$ converge almost surely. That is, associating the original random variables with a probability space $(\Omega, \mathscr{F}, P)$ involving a sample space Ω and the associated σ-algebra $\mathscr{F}$ of events, there exists an event $A \in \mathscr{F}$ with $P(A) = 1$ such that, for any sample point

$\omega \in A$, it holds that $\widehat{\boldsymbol{\theta}}_{n_k} \equiv \widehat{\boldsymbol{\theta}}_{n_k}(\omega) \to \boldsymbol{\theta}_0$ and $\rho(\mathcal{L}^*_{n_k}, \mathcal{L}) \equiv \rho(\mathcal{L}^*_{n_k}, \mathcal{L})(\omega) \to 0$ as $n_k \to \infty$ (i.e., pointwise convergence at each $\omega \in A$ along the subsequence $\{n_k\}$). Note that, given $\omega \in A$, estimates $\widehat{\boldsymbol{\theta}}_{n_k}(\omega)$ correspond to a single real sequence and there is a single sequence of bootstrap distributions $\mathcal{L}^*_{n_k}$, $n_k \geq 1$, for the bootstrap estimators $\sqrt{n_k}(\widehat{\boldsymbol{\theta}}^*_{n_k} - \widehat{\boldsymbol{\theta}}_{n_k}(\omega))$ induced by the bootstrap probability $\Pr_*$. To simplify the notation, we shall fix $\omega \in A$ and consider pointwise convergence at $\omega \in A$ as $n_k \to \infty$, suppressing the appearance of ω in the notation. Standard convergence in probability or distribution, though, with respect to the bootstrap probability $\Pr_*$ along the subsequence $\{n_k\}$, will be denoted as $\xrightarrow{d^*}$ and $\xrightarrow{p^*}$, respectively, for clarity. Hence along the subsequence $\{n_k\}$, it holds that $\widehat{\boldsymbol{\theta}}_{n_k} \to \boldsymbol{\theta}_0$ and $\sqrt{n_k}(\widehat{\boldsymbol{\theta}}^*_{n_k} - \widehat{\boldsymbol{\theta}}_{n_k}(\omega)) \xrightarrow{d^*} \boldsymbol{Z}_2$ for the multivariate normal random vector $\boldsymbol{Z}_2 \sim N(\boldsymbol{0}, \boldsymbol{V}_0)$ from Lemma 1.

We next define $\nabla(\boldsymbol{\theta}) = \partial\pi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, which is assumed to exist in a neighborhood of $\boldsymbol{\theta}_0$. Note that $\widehat{\boldsymbol{\theta}}^*_{n_k} \xrightarrow{p*} \boldsymbol{\theta}_0$ follows by $\widehat{\boldsymbol{\theta}}^*_{n_k} - \widehat{\boldsymbol{\theta}}_{n_k} \xrightarrow{p*} 0$ and $\widehat{\boldsymbol{\theta}}_{n_k} \to \boldsymbol{\theta}_0$. Consequently, for $\widehat{p}^*_{n_k} = \pi(\widehat{\boldsymbol{\theta}}^*_{n_k})$ and $\widehat{p}_{n_k} = \pi(\widehat{\boldsymbol{\theta}}_{n_k})$ based on the (continuously differentiable near $\boldsymbol{\theta}_0$) parametric function $\pi(\cdot)$ in (3), we use a Taylor expansion of $\pi(\widehat{\boldsymbol{\theta}}^*_{n_k})$ around $\widehat{\boldsymbol{\theta}}_{n_k}$ to obtain

$$\sqrt{n_k}(\widehat{p}^*_{n_k} - \widehat{p}_{n_k}) = \left[\nabla(\boldsymbol{c}^*_{n_k})\right]^t \sqrt{n_k}(\widehat{\boldsymbol{\theta}}^*_{n_k} - \widehat{\boldsymbol{\theta}}_{n_k}),$$

where $\nabla(\boldsymbol{c}^*_{n_k})$ is the gradient $\nabla(\boldsymbol{\theta})$ evaluated at $\boldsymbol{c}^*_{n_k} = \alpha^*_{n_k}\widehat{\boldsymbol{\theta}}_{n_k} + (1-\alpha^*_{n_k})\widehat{\boldsymbol{\theta}}^*_{n_k}$ for some $\alpha^*_{n_k} \in [0,1]$. Because $\widehat{\boldsymbol{\theta}}_{n_k} \to \boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}^*_{n_k} \xrightarrow{p^*} \widehat{\boldsymbol{\theta}}_{n_k}$, we have

$$\begin{aligned}
\|\boldsymbol{c}^*_{n_k} - \boldsymbol{\theta}_0\| &= \|\alpha^*_{n_k}(\widehat{\boldsymbol{\theta}}_{n_k} - \boldsymbol{\theta}_0) + (1-\alpha^*_{n_k})(\widehat{\boldsymbol{\theta}}^*_{n_k} - \boldsymbol{\theta}_0)\| \\
&\leq \|\alpha^*_{n_k}(\widehat{\boldsymbol{\theta}}_{n_k} - \boldsymbol{\theta}_0)\| + \|(1-\alpha^*_{n_k})(\widehat{\boldsymbol{\theta}}^*_{n_k} - \widehat{\boldsymbol{\theta}}_{n_k})\| + \|((1-\alpha^*_{n_k})(\widehat{\boldsymbol{\theta}}_{n_k} - \boldsymbol{\theta}_0)\| \\
&\leq 2\|\widehat{\boldsymbol{\theta}}_{n_k} - \boldsymbol{\theta}_0\| + \|\widehat{\boldsymbol{\theta}}^*_{n_k} - \widehat{\boldsymbol{\theta}}_{n_k}\| \xrightarrow{p^*} 0
\end{aligned}$$

as $n_k \to \infty$. Because $\nabla(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$, the continuous mapping theorem then gives

$$\nabla(\boldsymbol{c}^*_{n_k}) \xrightarrow{p^*} \nabla(\boldsymbol{\theta}_0) \equiv \nabla_0$$

and then Slutsky's theorem yields $\sqrt{n_k}(\widehat{p}^*_{n_k} - \widehat{p}_{n_k}) \xrightarrow{d^*} \nabla_0^t \boldsymbol{Z}_2$. Because the random variable $\nabla_0^t \boldsymbol{Z}_2$ is continuous with cdf $\Phi_{\text{nor}}(z/\sqrt{v_0})$, $z \in \mathbb{R}$, Polya's theorem then implies uniform convergence of cdfs as

$$\sup_{z\in\mathbb{R}} \left|\Pr_*(\sqrt{n_k}(\widehat{p}^*_{n_k} - \widehat{p}_{n_k}) \le z) - \Phi_{\text{nor}}(z/\sqrt{v_0})\right| \to 0$$

as $n_k \to \infty$. Because the above distance between distributions converges almost surely to zero along the subsequence $\{n_k\} \subset \{n_j\}$ and because the subsequence $\{n_j\}$ was arbitrary, we have shown that every subsequence contains a further subsequence where this distributional distance converges to zero almost surely; the probabilistic convergence in Part 2 of Lemma 2 then follows.

For the GPQ-bootstrap method in Part 3 of Lemma 2, we define $h(z, x, y) = \Phi\left[z(1 + y/\widehat{\sigma}_{n_k}) + x/\widehat{\sigma}_{n_k}\right]$, $\widehat{z}_w = [\log(t_w) - \widehat{\mu}_{n_k}]/\widehat{\sigma}_{n_k}$ and $\widehat{z}_c = [\log(t_c) - \widehat{\mu}_{n_k}]/\widehat{\sigma}_{n_k}$. Here $\Phi(\cdot) = F(\cdot; 0, 1)$ is the standard cdf of the log-location-scale distribution with derivative $\Phi'(\cdot) \equiv \phi(\cdot)$ on $\mathbb{R}$, and $\widehat{\mu}_{n_k}$, $\widehat{\sigma}_{n_k}$ are the consistent estimators of μ and σ. Then we define $g_n(x, y) = \left[h(\widehat{z}_w, x, y) - h(\widehat{z}_c, x, y)\right]/[1 - h(\widehat{z}_c, x, y)]$, so that $\widehat{p}^{**}_{n_k} - \widehat{p}_{n_k} = g_n(\widehat{\mu}^*_{n_k} - \widehat{\mu}_{n_k}, \widehat{\sigma}^*_{n_k} - \widehat{\sigma}_{n_k}) - g_n(0, 0)$. We use a Taylor expansion of $g_n(\widehat{\mu}^*_{n_k} - \widehat{\mu}_{n_k}, \widehat{\sigma}^*_{n_k} - \widehat{\sigma}_{n_k})$ at $(0, 0)$ to obtain

$$\sqrt{n_k}(\widehat{p}^{**}_{n_k} - \widehat{p}_{n_k}) = \left[(\partial g_n/\partial x, \partial g_n/\partial y)|_{x=0,y=0} + R^*_n\right] \sqrt{n_k} \left(\widehat{\mu}^*_{n_k} - \widehat{\mu}_{n_k}, \widehat{\sigma}^*_{n_k} - \widehat{\sigma}_{n_k}\right)^t,$$

where $R_n \xrightarrow{p^*} 0$ by the differentiability of $g_n(x, y)$ at $(0, 0)$ combined with $\sqrt{n}(\widehat{\mu}^*_{n_k} - \widehat{\mu}_{n_k}, \widehat{\sigma}^*_{n_k} - \widehat{\sigma}_{n_k}) \xrightarrow{p^*} 0$ and $(\widehat{\mu}_{n_k}, \widehat{\sigma}_{n_k}) \to (\mu_0, \sigma_0)$.

Because $(\widehat{\mu}_{n_k}, \widehat{\sigma}_{n_k}) \to (\mu_0, \sigma_0)$ and $R_n \xrightarrow{p} 0$ as $n_k \to \infty$, we have

$$\begin{aligned}\left(\frac{\partial g_n}{\partial x}, \frac{\partial g_n}{\partial y}\right)^t \Bigg|_{x=0,y=0} &\xrightarrow{p^*} \begin{pmatrix} \frac{1}{\sigma[1-\phi(z_c)]^2}\left\{\phi(z_c)[1 - \Phi(z_w)] - \phi(z_w)[1 - \Phi(z_c)]\right\} \\ \frac{1}{\sigma[1-\phi(z_c)]^2}\left\{z_c\phi(z_c)[1 - \Phi(z_w)] - z_w\phi(z_w)[1 - \Phi(z_c)]\right\} \end{pmatrix} \\ &= \left(\frac{\partial \pi(\mu, \sigma)}{\partial \mu}, \frac{\partial \pi(\mu, \sigma)}{\partial \sigma}\right)^t \Bigg|_{\mu=\mu_0,\sigma=\sigma_0} = \nabla_0,\end{aligned}$$

where we define $z_c = [\log(t_c) - \mu_0]/\sigma_0$, $z_w = [\log(t_w) - \mu_0]/\sigma_0$ and $\phi(\cdot)$ is the pdf/derivative of $\Phi(\cdot)$. The rest of the proof follows by Slutsky's theorem in the same manner as the proof for the direct-bootstrap method. □

Lemma 3. *Let Z_0 and Z_1 denote independent standard normal random variables. Under the conditions for Theorem 1 conditions with true parameters $\boldsymbol{\theta}_0$, the following (1)-(3) hold as $n \to \infty$*

(1)

$$\frac{Y_n - (n - r_n)\widehat{p}_n}{\sqrt{(n - r_n)\widehat{p}_n(1 - \widehat{p}_n)}} \xrightarrow{d} Z_0 + \sqrt{v_1} Z_1,$$

where $v_1 = v_0[1 - F(t_c; \boldsymbol{\theta}_0)]/[p_0(1 - p_0)]$ for $p_0 = \pi(\boldsymbol{\theta}_0)$ and v_0 is from Lemma 2.

(2) Based on the bootstrap version $\widehat{p}_n^* = \pi(\widehat{\boldsymbol{\theta}}_n^*)$ of $\widehat{p}_n = \pi(\widehat{\boldsymbol{\theta}}_n)$, let the random variable Y_n^* be defined as $Y_n^* | \widehat{p}_n^* \sim \text{Binomial}(n - r_n, \widehat{p}_n^*)$, where r_n is the number of events in the given sample. Then it holds that

$$\sup_{z \in \mathbb{R}} \left| \Pr_* \left[\frac{Y_n^* - (n - r_n)\widehat{p}_n}{\sqrt{(n - r_n)\widehat{p}_n(1 - \widehat{p}_n)}} \leq z \right] - \Pr(Z_0 + \sqrt{v_1} Z_1 \leq z) \right| \xrightarrow{p} 0.$$

(3) Based on the approximate GPQ-based bootstrap version $\widehat{p}_n^{**} = \pi(\widehat{\mu}_n^{**}, \widehat{\sigma}_n^{**})$ of $\widehat{p}_n = \pi(\widehat{\mu}_n, \widehat{\sigma}_n)$ and a random variable Y_n^{**} defined as $Y_n^{**} | \widehat{p}_n^{**} \sim \text{Binomial}(n - r_n, \widehat{p}_n^{**})$, it holds that

$$\sup_{z \in \mathbb{R}} \left| \Pr_* \left[\frac{Y_n^{**} - (n - r_n)\widehat{p}_n}{\sqrt{(n - r_n)\widehat{p}_n(1 - \widehat{p}_n)}} \leq z \right] - \Pr(Z_0 + \sqrt{v_1} Z_1 \leq z) \right| \xrightarrow{p} 0.$$

Proof. Fix $z \in \mathbb{R}$, based on the censored sample $\boldsymbol{D}_n$, define an event $M_n = M_{1n} \cap M_{2n}$, where $M_{1n} = \{r_n < n\}$ and $M_{2n} = \{0 < \widehat{p}_n < 1\}$. As $n \to \infty$, note that $\Pr(M_{2n}^c) \to 0$ by Lemma 2 Part 1 with $p_0 \in (0, 1)$ (cf. Theorem 1), while $\Pr(M_{1n}^c) = \Pr(r_n = n) = [F(t_c; \boldsymbol{\theta}_0)]^n \to 0$ (i.e., $F(t_c; \boldsymbol{\theta}_0) \in (0, 1)$ under Theorem 1 conditions). Hence, it follows that $\Pr(M_n) \to 1$ as $n \to \infty$ and the predictive root $[Y_n - (n - r_n)\widehat{p}_n] / \sqrt{(n - r_n)\widehat{p}_n(1 - \widehat{p}_n)}$ is well defined when the event M_n holds. Hence, for fixed $z \in \mathbb{R}$, we may write

$$\left| \Pr \left[\frac{Y_n - (n - r_n)\widehat{p}_n}{\sqrt{(n - r_n)\widehat{p}_n(1 - \widehat{p}_n)}} \leq z \right] - \tau_n \right| \leq \Pr(M_n^c) \to 0 \tag{C.1}$$

for

$$\begin{aligned}\tau_n &\equiv \Pr\left[M_n, \frac{Y_n - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \leq z\right] \\ &= \Pr\left[M_n, \frac{Y_n - (n-r_n)p_0}{\sqrt{(n-r_n)p_0(1-p_0)}} \leq z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{p_0(1-p_0)}} + \frac{\sqrt{n-r_n}(\widehat{p}_n - p_0)}{\sqrt{p_0(1-p_0)}}\right].\end{aligned}$$

Conditioning on the censored data $\boldsymbol{D}_n$, we further write a conditional probability version of τ_n as

$$\begin{aligned}&\Pr\left[M_n, \frac{Y_n - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \leq z\middle|\boldsymbol{D}_n\right] \\ =&\mathrm{I}(M_n)\Pr\left[\frac{Y_n - (n-r_n)p_0}{\sqrt{(n-r_n)p_0(1-p_0)}} \leq z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{p_0(1-p_0)}} + \frac{\sqrt{n-r_n}(\widehat{p}_n - p_0)}{\sqrt{p_0(1-p_0)}}\middle|\boldsymbol{D}_n\right] \\ =&\mathrm{I}(M_n)\Phi_{\mathrm{nor}}\left[z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{p_0(1-p_0)}} + \frac{\sqrt{n-r_n}(\widehat{p}_n - p_0)}{\sqrt{p_0(1-p_0)}}\right] + \mathrm{I}(M_n)R_n\end{aligned}$$

where $\Phi_{\mathrm{nor}}(\cdot)$ denotes a standard normal cdf, $\mathrm{I}(\cdot)$ denotes the indicator function, and R_n is a remainder that satisfies

$$|R_n| \leq \frac{1}{\sqrt{(n-r_n)p_0(1-p_0)}}$$

by the Berry-Esseen theorem applied to $(n-r_n)$ independent Bernoulli(p_0) random variables. As $n \to \infty$, note that $(n-r_n)/n \xrightarrow{p} 1 - F(t_c;\boldsymbol{\theta}_0)$ by the weak law of large numbers, so that $R_n \xrightarrow{p} 0$ follows as well as

$$\frac{\sqrt{n-r_n}(\widehat{p}_n - p_0)}{\sqrt{p_0(1-p_0)}} = \frac{[(n-r_n)/n]^{1/2}}{\sqrt{p_0(1-p_0)}}\sqrt{n}(\widehat{p}_n - p_0) \xrightarrow{d} \frac{[1-F(t_c;\boldsymbol{\theta}_0)]^{1/2}}{\sqrt{p_0(1-p_0)}}\sqrt{v_0}Z_1 = \sqrt{v_1}Z_1$$

by Slutsky's theorem with Lemma 2 Part 1. Along with $\mathrm{I}(M_n) \xrightarrow{p} 1$, $\widehat{p}_n \xrightarrow{p} p_0$ and the continuity of $\Phi_{\mathrm{nor}}(\cdot)$, the continuous mapping theorem then yields

$$\Pr\left[M_n, \frac{Y_n - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \leq z\middle|\boldsymbol{D}_n\right] \xrightarrow{d} \Phi_{\mathrm{nor}}(z + \sqrt{v_1}Z_1).$$

Because this conditional probability is bounded by 1 and hence uniformly integrable, its convergence in distribution also implies convergence of its expectation: as $n \to \infty$,

$$\tau_n = \Pr\left[M_n, \frac{Y_n - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \le z\right] = \mathrm{E}\left\{\Pr\left[M_n, \frac{Y_n - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \le z \middle| \boldsymbol{D}_n\right]\right\}$$
$$\to \mathrm{E}[\Phi_{\mathrm{nor}}(z + \sqrt{v_1}Z_1)].$$

Consequently, by the above with (C.1), we have that

$$\Pr\left[\frac{Y_n - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \le z\right] \to \mathrm{E}[\Phi_{\mathrm{nor}}(z + \sqrt{v_1}Z_1)] = \Pr(Z_0 + \sqrt{v_1}Z_1 \le z),$$

where $\mathrm{E}[\Phi_{\mathrm{nor}}(z + \sqrt{v_1}Z_1)] = \Pr(Z_0 + \sqrt{v}Z_1 \le z)$ follows for iid standard normal variables Z_0, Z_1. Because $z \in \mathbb{R}$ is arbitrary, we have that the cdf of the predictive root $[Y_n - (n - r_n)\widehat{p}_n]/\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}$ converges to the cdf of $Z_0 + \sqrt{v_1}Z_1$ for any z and hence $Y_n - (n-r_n)\widehat{p}_n/\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)} \xrightarrow{d} Z_0 + \sqrt{v_1}Z_1$.

The proof of Lemma 3 Part 2 closely follows the argument for Lemma 2 Part 2. Let $\{n_j\} \subset \{n\}$ be an arbitrary subsequence of indices. Because $\widehat{p}_n \xrightarrow{p} p_0 > 0$ holds by Lemma 2 along with the facts that $(n - r_n)/n \xrightarrow{p} 1 - F(t_c; \boldsymbol{\theta}_0)$ by the weak law of large numbers, while the bootstrap distribution of $\sqrt{n_k}(\widehat{p}^*_{n_k} - \widehat{p}_{n_k})$ converges in probability under Lemma 2, we may extract a further subsequence $\{n_k\} \subset \{n_j\}$ along which $\widehat{p}_{n_k} \to p_0$, $(n_k - r_{n_k})/n_k \to 1 - F(t_c; \boldsymbol{\theta}_0) > 0$ and $\sqrt{n_k}(\widehat{p}^*_{n_k} - \widehat{p}_{n_k}) \xrightarrow{d^*} \sqrt{v_0}Z_1$ converge almost surely. As in the proof of Lemma 2, we again consider the subsequence $\{n_k\}$ as $n_k \to \infty$ for a fixed point $\omega \in A$ defined by an event A with $\Pr(A) = 1$ where the above-mentioned almost sure convergence holds. Fix $z \in \mathbb{R}$. Then for large n_k, where $n_k > r_{n_k}$ is then guaranteed, the conditional bootstrap distribution $Y^*_{n_k}|\widehat{p}^*_{n_k}$ is $\mathrm{Binomial}(n_k - r_{n_k}, \widehat{p}^*_{n_k})$ so that, by the Berry-Esseen theorem applied to the sum of $n_k - r_{n_k}$

iid $\mathrm{Bernoulli}(\widehat{p}^*_{n_k})$ variables, we have

$$\begin{aligned}
& \Pr_* \left[\frac{Y^*_{n_k} - (n_k - r_{n_k})\widehat{p}_{n_k}}{\sqrt{(n_k - r_{n_k})\widehat{p}_{n_k}(1-\widehat{p}_{n_k})}} \le z \middle| \widehat{p}^*_{n_k} \right] \\
= \; & \Pr_* \left[\frac{Y^*_{n_k} - (n_k - r_{n_k})\widehat{p}^*_{n_k}}{\sqrt{(n_k - r_{n_k})\widehat{p}^*_{n_k}(1-\widehat{p}^*_{n_k})}} \le z \frac{\sqrt{\widehat{p}_{n_k}(1-\widehat{p}_{n_k})}}{\sqrt{\widehat{p}^*_{n_k}(1-\widehat{p}^*_{n_k})}} + \frac{\sqrt{n_k - r_{n_k}}(\widehat{p}_{n_k} - \widehat{p}^*_{n_k})}{\sqrt{\widehat{p}^*_{n_k}(1-\widehat{p}^*_{n_k})}} \middle| \widehat{p}^*_{n_k} \right] \\
= \; & \Phi_{\mathrm{nor}} \left[z \frac{\sqrt{\widehat{p}_{n_k}(1-\widehat{p}_{n_k})}}{\sqrt{\widehat{p}^*_{n_k}(1-\widehat{p}^*_{n_k})}} + \frac{\sqrt{n_k - r_{n_k}}(\widehat{p}_{n_k} - \widehat{p}^*_{n_k})}{\sqrt{\widehat{p}^*_{n_k}(1-\widehat{p}^*_{n_k})}} \right] + R^*_{n_k}
\end{aligned}$$

where $\Phi_{\mathrm{nor}}(\cdot)$ denotes the standard normal cdf and $R^*_{n_k}$ is a remainder bounded by

$$|R^*_{n_k}| \le \frac{1}{\sqrt{n_k - r_{n_k}}} \frac{1}{\sqrt{\widehat{p}^*_{n_k}(1-\widehat{p}^*_{n_k})}}.$$

Note that we are technically assuming that $0 < \widehat{p}^*_{n_k} < 1$ in the conditioning of the bootstrap conditional probability above to simplify the argument, which is asymptotically valid though as $\Pr_*(0 < \widehat{p}^*_{n_k} < 1) \to 1$ (cf. the next line). Because $(n_k - r_{n_r})/n_k \to 1 - F_0(t_c)$, $\widehat{p}_{n_k} \to p_0 \in (0,1)$ and $-\sqrt{n_k}(\widehat{p}^*_{n_k} - \widehat{p}_{n_k}) \xrightarrow{d*} \sqrt{v_0} Z_1$ (with the latter two properties also implying that $\widehat{p}^*_{n_k} \xrightarrow{p*} p_0$) as $n_k \to \infty$, it follows by Slutsky's theorem that $R^*_{n_k} \xrightarrow{p*} 0$ and

$$z \frac{\sqrt{\widehat{p}_{n_k}(1-\widehat{p}_{n_k})}}{\sqrt{\widehat{p}^*_{n_k}(1-\widehat{p}^*_{n_k})}} + \frac{\sqrt{n_k - r_{n_k}}(\widehat{p}_{n_k} - \widehat{p}^*_{n_k})}{\sqrt{\widehat{p}^*_{n_k}(1-\widehat{p}^*_{n_k})}} \xrightarrow{d*} z + \frac{[1-F_0(t_c)]^{1/2}}{\sqrt{p_0(1-p_0)}} \sqrt{v_0} Z_1 = z + \sqrt{v_1} Z_1,$$

so that the continuous mapping theorem gives

$$\Pr_* \left[\frac{Y^*_{n_k} - (n_k - r_{n_k})\widehat{p}_{n_k}}{\sqrt{(n_k - r_{n_k})\widehat{p}_{n_k}(1-\widehat{p}_{n_k})}} \le z \middle| \widehat{p}^*_{n_k} \right] \xrightarrow{d*} \Phi_{\mathrm{nor}}(z + \sqrt{v_1} Z_1)$$

by the continuity of $\Phi_{\mathrm{nor}}(\cdot)$. Because the above bootstrap conditional probability $\Pr_*(\cdot|\widehat{p}^*_{n_k})$ is bounded by 1 and converges in distribution (under bootstrap probability $\Pr_*$ as $n_k \to \infty$), its bootstrap expectation E_* (i.e., under $\Pr_*$) also converges

$$\begin{aligned}
\Pr_* \left[\frac{Y^*_{n_k} - (n_k - r_{n_k})\widehat{p}_{n_k}}{\sqrt{(n_k - r_{n_k})\widehat{p}_{n_k}(1-\widehat{p}_{n_k})}} \le z \right] &= \mathrm{E}_* \left\{ \Pr_* \left[\frac{Y^*_{n_k} - (n_k - r_{n_k})\widehat{p}_{n_k}}{\sqrt{(n_k - r_{n_k})\widehat{p}_{n_k}(1-\widehat{p}_{n_k})}} \le z \middle| \widehat{p}^*_{n_k} \right] \right\} \\
&\to \mathrm{E}\Phi_{\mathrm{nor}}(z + \sqrt{v_1} Z_1) = \Pr(Z_0 + \sqrt{v_1} Z_1 \le z)
\end{aligned}$$

as $n_k \to \infty$. Because $z \in \mathbb{R}$ was arbitrary and the cdf of $Z_0 + \sqrt{v_1}Z_1$ is continuous, we have

$$\sup_{z\in\mathbb{R}} \left| \Pr_* \left[\frac{Y^*_{n_k} - (n_k - r_k)\widehat{p}_{n_k}}{\sqrt{(n_k - r_k)\widehat{p}_{n_k}(1-\widehat{p}_{n_k})}} \leq z \right] - \Pr\left(Z_0 + \sqrt{v_1}Z_1 \leq z\right) \right| \to 0$$

(pointwise/almost surely) as $n_k \to \infty$. As this last convergence to zero holds almost surely along the subsequence $\{n_k\} \subset \{n_j\}$ and, as the subsequence $\{n_j\}$ was arbitrary, we have shown that this convergence to zero must hold in probability (along n) and Lemma 3 Part 2 follows. Finally, the proof of Lemma 3 Part 3 follows by substituting $(Y^{**}_n, \widehat{p}^{**}_n)$ for $(Y^*_n, \widehat{p}^*_n)$ in the proof of Lemma 3 Part 2. □

Proof of Theorem 1

Proof. We define $\Delta_n(y) = \text{pbinom}(y, n - r_n, \widehat{p}_n) - \text{pbinom}(y, n - r_n, p_0)$. To prove Theorem 1 Part 1, without loss of generality, we often assume that $0 < \widehat{p}_n < 1$ and $r_n < n$, as $\widehat{p}_n \xrightarrow{p} p_0 \in (0,1)$ by Lemma 2 and $\Pr(r_n = n) = [F(t_c; \boldsymbol{\theta}_0)]^n \to 0$ by $F(t_c; \boldsymbol{\theta}_0) \in (0,1)$. Using the Berry-Esseen theorem we have,

$$\begin{aligned}
\sup_{y\in\mathbb{R}} |\Delta_n(y)| &= \sup_{y\in\mathbb{R}} \left| \Phi_{\text{nor}} \left[\frac{y - (n-r_n)p_0}{\sqrt{(n-r_n)p_0(1-p_0)}} \right] - \Phi_{\text{nor}} \left[\frac{y - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \right] \right| + R_n \\
&= \sup_{z\in\mathbb{R}} |\Phi_{\text{nor}}(z) - \Phi_{\text{nor}}(zA_n + B_n)| + R_n,
\end{aligned}$$

where $A_n \equiv \sqrt{p_0(1-p_0)}/\sqrt{\widehat{p}_n(1-\widehat{p}_n)}$, $B_n \equiv -(n-r_n)(\widehat{p}_n - p_0)/\sqrt{\widehat{p}_n(1-\widehat{p}_n)}$ and $|R_n| \leq 2\text{I}(\widehat{p}_n \in \{0,1\}) + 2\text{I}(r_n = n) + [p_0(1-p_0)]^{-1/2}(n-r_n)^{-1/2}\text{I}(r_n < n) + [\widehat{p}_n(1-\widehat{p}_n))]^{-1/2}(n - r_n)^{-1/2}\text{I}(0 < \widehat{p}_n < 1, r_n < n)$ for $\text{I}(\cdot)$ denoting an indicator function and $\Phi_{\text{nor}}(\cdot)$ denoting the standard normal cdf. Because $\widehat{p}_n \xrightarrow{p} p_0 \in (0,1)$ by Lemma 2 and $(n-r_n)/n \xrightarrow{p} 1 - F(t_c; \boldsymbol{\theta}_0) \in (0,1)$ by the weak law of large numbers, we have $|R_n| \xrightarrow{p} 0$.

Note $m(a,b) \equiv \sup_{z\in\mathbb{R}} |\Phi_{\text{nor}}(z) - \Phi_{\text{nor}}(az + b)|$ is continuous as a function of $(a,b) \in (0,\infty) \times \mathbb{R}$. By Lemma 1, $(A_n, B_n) \xrightarrow{d} (1, \sqrt{v_1}Z_1)$ for $Z_1 \sim N(0,1)$. By the continuous mapping theorem and Slutsky's theorem, we then have

$$\sup_{y\in\mathbb{R}} |\Delta_n(y)| \xrightarrow{d} m(1, \sqrt{v_1}Z_1) = \sup_{z\in\mathbb{R}} |\Phi_{\text{nor}}(z) - \Phi_{\text{nor}}(z + \sqrt{v_1}Z_1)| = 1 - 2\Phi_{\text{nor}}\left(-\frac{\sqrt{v_1}}{2}|Z_1| \right);$$

the latter supremum is determined at an argument value of $z = \text{sign}(-Z_1)\sqrt{v_1}|Z_1|/2$.

For Theorem 1 Part 2, we first show that the plug-in method produces an upper prediction bound $\tilde{Y}^{PL}_{n,1-\alpha}$ such that

$$\frac{\tilde{Y}^{PL}_{n,1-\alpha} - (n-r)\widehat{p}_n}{\sqrt{(n-r)\widehat{p}_n(1-\widehat{p}_n)}} \xrightarrow{p} \Phi^{-1}_{\text{nor}}(1-\alpha), \tag{C.2}$$

where $\Phi^{-1}_{\text{nor}}(1-\alpha)$ denotes the $100(1-\alpha)\%$ quantile of a standard normal variable Z_0 with cdf $\Phi_{\text{nor}}(\cdot)$, and we write $r \equiv r_n$. This follows because the plug-in method uses the $1-\alpha$ quantile of a Binomial$(n-r, \widehat{p}_n)$ random variable $Y_{0,n}$ for calibration so that $\tilde{Y}^{PL}_{n,1-\alpha} \approx \Phi^{-1}_{\text{nor}}(1-\alpha)\sqrt{(n-r)\widehat{p}_n(1-\widehat{p}_n)} + (n-r)\widehat{p}_n$. More formally, by the Central Limit Theorem (or, by the Berry-Esseen theorem) applied to the sum of $(n-r)$ iid Bernoulli$(\widehat{p}_n)$ variables, we find

$$\sup_{z\in\mathbb{R}} \left| \Pr_*\left(Y_{0,n} \le z\right) - \Phi_{\text{nor}}\left[z\sqrt{(n-r)\widehat{p}_n(1-\widehat{p}_n)} + (n-r)\widehat{p}_n\right] \right| \le$$
$$2\text{I}(\widehat{p}_n \in \{0,1\}) + 2\text{I}(r=n) + \frac{1}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}\frac{1}{\sqrt{n-r}}\text{I}(0<\widehat{p}_n<1, r<n) \xrightarrow{p} 0$$

by $(n-r)/n \xrightarrow{p} 1 - F(t_c;\boldsymbol{\theta}_0) > 0$ and $\widehat{p}_n \xrightarrow{p} p_0$ again. This implies (C.2) as $\tilde{Y}^{PL}_{n,1-\alpha} = \inf\{z \in \mathbb{R} : \Pr_*(Y_{0,n} \le z) \ge 1-\alpha\}$.

By (C.2) and Lemma 3, we then have

$$\frac{Y_n - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} - \frac{\tilde{Y}^{PL}_{n,1-\alpha} - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \xrightarrow{d} Z_0 + \sqrt{v_1}Z_1 - \Phi^{-1}_{\text{nor}}(1-\alpha)$$

by Slutsky's theorem. By the last line, it follows immediately that

$$\begin{aligned}\Pr(Y_n \le \tilde{Y}^{PL}_{n,1-\alpha}) &= \Pr\left[\frac{Y_n - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \le \frac{\tilde{Y}^{PL}_{n,1-\alpha} - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}}\right] \\ &= \Pr\left[Z_0 + \sqrt{v_1}Z_1 \le \Phi^{-1}_{\text{nor}}(1-\alpha)\right] \quad \text{as} \quad n\to\infty.\end{aligned}$$

Note that v_1 here is based on its occurrence in Lemma 3 and that, as Z_0 and Z_1 are i.i.d. standard normal, we may write

$$\begin{aligned}\Pr\left[Z_0 + \sqrt{v_1}Z_1 \le \Phi^{-1}_{\text{nor}}(1-\alpha)\right] &= \text{E}\Phi_{\text{nor}}[\Phi^{-1}_{\text{nor}}(1-\alpha) - \sqrt{v_1}Z_1] \\ &= \int_{-\infty}^{\infty} \Phi_{\text{nor}}[\Phi^{-1}_{\text{nor}}(1-\alpha) + \sqrt{v_1}z]\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz \equiv \Lambda_{1-\alpha}(v_1).\end{aligned}$$

Note that the probability above is 0.5 when $\alpha = 0.5$.

As a function of $v_1 > 0$ (with fixed $\alpha \in (0,1) \setminus \{0.5\}$), $\Lambda_{1-\alpha}(v_1)$ is differentiable with derivative

$$\begin{aligned}\Lambda'_{1-\alpha}(v_1) &= 0.5 v_1^{-3/2} \int_{-\infty}^{\infty} \phi_{\rm nor}[\Phi_{\rm nor}^{-1}(1-\alpha) + \sqrt{v_1} z] z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &\equiv 0.5 v_1^{-3/2} \int_{0}^{\infty} [t_\alpha(z; v_1) - t_\alpha(-z; v_1)] z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz,\end{aligned}$$

for

$$t_\alpha(z; v_1) \equiv \phi_{\rm nor}[\Phi_{\rm nor}^{-1}(1-\alpha) + \sqrt{v_1} z], \quad z \in \mathbb{R}.$$

For fixed $z > 0$ and $v_1 > 0$, it holds that $t_\alpha(z; v_1)/t_\alpha(-z; v_1) < 1$ if $\alpha \in (0, 0.5)$ (which may be checked with simple algebra), and that $t_\alpha(z; v_1)/t_\alpha(-z; v_1) > 1$ if $\alpha \in (0.5, 1)$. Consequently, it follows that the derivative $\Lambda'_{1-\alpha}(v_1) < 0$ for all $v_1 > 0$ if $\alpha \in (0, 0.5)$, while $\Lambda'_{1-\alpha}(v_1) > 0$ for all $v_1 > 0$ if $\alpha \in (0.5, 1)$; that is, $\Lambda_{1-\alpha}(v_1)$ is decreasing on $v_1 \in (0, \infty)$ if $\alpha \in (0, 0.5)$, and increasing on $v_1 \in (0, \infty)$ if $\alpha \in (0.5, 1)$. Further, as $\Phi_{\rm nor}(\Phi_{\rm nor}^{-1}(1-\alpha) + \sqrt{v_1} z)$ is bounded by 1 and converges, for each fixed real $z \neq 0$, to the indicator function $\mathrm{I}(z > 0)$ as $v \to \infty$, the dominated convergence theorem gives

$$\lim_{v_1 \to \infty} \Lambda_{1-\alpha}(v_1) = \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.5.$$

Note as well that when $\alpha \in (0, 0.5)$, we have for any $v_1 > 0$ that

$$\Lambda_{1-\alpha}(v_1) < \lim_{v_1 \downarrow 0} \Lambda_{1-\alpha}(v_1) = \Lambda_{1-\alpha}(0) = 1 - \alpha,$$

as $\Lambda_{1-\alpha}(v_1)$ is decreasing; when $\alpha \in (0, 0.5)$, we have instead that $\Lambda_{1-\alpha}(v_1) > 1 - \alpha$ when $\alpha \in (0.5, 1)$. □

Proof of Theorem 2

Proof. For $\Psi_n^* \equiv [Y_n^* - (n - r_n^*)\widehat{p}_n^*]/\sqrt{(n - r_n^*)\widehat{p}_n^*(1 - \widehat{p}_n^*)}$, we first want to show that

$$\sup_{y \in \mathbb{R}} |\Pr_*(\Psi_n^* \leq y) - \Pr(Z_0 + \sqrt{v_1} Z_1 \leq y)| \xrightarrow{p} 0, \text{as } n \to \infty,$$

where $Y_n^* \sim \text{Binomial}(n - r_n^*, \widehat{p}_n)$ conditional on $(n - r_n^*)$, where r_n^* is the number of events in the bootstrap sample and $\widehat{p}_n^*$ is the bootstrap version of $\widehat{p}_n$. In the bootstrap world, we have $(n - r_n^*)/n = 1 - \sum_{i=1}^n \text{I}(T_i^* \leq t_c)/n$, where under bootstrap expectation

$$\text{E}_* \left(\frac{n - r_n^*}{n} \right) = 1 - \Pr_*(T_1^* \leq t_c) \xrightarrow{p} 1 - F(t_c, \boldsymbol{\theta}_0) > 0,$$

by the assumption that $\Pr_*(T_1^* \leq t_c)$ is consistent for $F(t_c; \boldsymbol{\theta}_0)$ (e.g., $\Pr_*(T_1^* \leq t_c) = F(t_c; \widehat{\boldsymbol{\theta}}_n)$), and likewise

$$\begin{aligned} \text{Var}_* \left(\frac{n - r_n^*}{n} \right) &= \text{Var}_* \left[\frac{1}{n} \sum_{i=1}^n \text{I}(T_i^* \leq t_c) \right] \\ &= \frac{1}{n} \Pr_* \left(T_1^* \leq t_c \right) \Pr_* \left(T_1^* > t_c \right) \leq \frac{1}{n} \xrightarrow{p} 0, \text{ as } n \to \infty. \end{aligned}$$

Hence, the convergence of these two bootstrap moments implies that $(n - r_n^*)/n$ converges in bootstrap probability or $(n - r_n^*)/ \xrightarrow{p*} 1 - F(t_c; \boldsymbol{\theta}_0)$ (in probability); that is, for any subsequence $\{n_j\} \subset \{n\}$, there exists a further subsequence $\{n_k\} \subset \{n_j\}$ where with probability 1,

$$\Pr_* \left[\left| \frac{n_k - r_{n_k}^*}{n_k} - 1 + F(t_c; \boldsymbol{\theta}_0) \right| > \epsilon \right] \to 0$$

holds as $n_k \to \infty$ for each given $\epsilon > 0$. Choose a subsequence $\{n_k\}$ where together $(n_k - r_{n_k}^*)/n_k \xrightarrow{p^*} 1 - F(t_c; \boldsymbol{\theta}_0) \in (0, 1)$ (by the above) and $\sqrt{n_k - r_{n_k}}(\widehat{p}_{n_k}^* - \widehat{p}_{n_k})/\sqrt{\widehat{p}_{n_k}(1 - \widehat{p}_{n_k})} \xrightarrow{d^*} \sqrt{v_1} Z_1$ (by Lemma 3) and $\widehat{p}_{n_k} \to p_0 \in (0, 1)$ and $(n_k - r_{n_k})/n_k \to 1 - F(t_c; \boldsymbol{\theta}_0)$ hold as $n_k \to \infty$ with probability 1. Now conditional on $(n_k - r_{n_k}^*), \widehat{p}_{n_k}^*$ (here, without a loss of generality, assuming $n - r_{n_k}^* > 0$, $0 < \widehat{p}_{n_k}^* < 1$) and for fixed $y \in \mathbb{R}$, we can write

$$\begin{aligned} &\Pr_*(\Psi_{n_k}^* \leq y | n_k - r_{n_k}^*, \widehat{p}_{n_k}^*) \\ =& \Pr_* \left[\frac{Y_{n_k}^* - (n_k - r_{n_k}^*)\widehat{p}_{n_k}}{\sqrt{(n_k - r_{n_k}^*)\widehat{p}_{n_k}(1 - \widehat{p}_{n_k})}} \leq y \frac{\sqrt{(n_k - r_{n_k}^*)\widehat{p}_{n_k}^*(1 - \widehat{p}_{n_k}^*)}}{\sqrt{(n_k - r_{n_k}^*)\widehat{p}_{n_k}(1 - \widehat{p}_{n_k})}} - \frac{\sqrt{n_k - r_{n_k}^*}(\widehat{p}_{n_k}^* - \widehat{p}_{n_k})}{\sqrt{\widehat{p}_{n_k}(1 - \widehat{p}_{n_k})}} \middle| n_k - r_{n_k}^*, \widehat{p}_{n_k}^* \right] \\ =& \Phi_{\text{nor}} \left[y \frac{\sqrt{(n_k - r_{n_k}^*)\widehat{p}_{n_k}^*(1 - \widehat{p}_{n_k}^*)}}{\sqrt{(n_k - r_{n_k}^*)\widehat{p}_{n_k}(1 - \widehat{p}_{n_k})}} - \frac{\sqrt{n_k - r_{n_k}^*}(\widehat{p}_{n_k}^* - \widehat{p}_{n_k})}{\sqrt{\widehat{p}_{n_k}(1 - \widehat{p}_{n_k})}} \right] + R_{n_k}^*, \end{aligned}$$

by the Berry-Esseen theorem, where the remainder $R^*_{n_k}$ is bounded by

$$\left|R^*_{n_k}\right| \leq 2\mathrm{I}(\widehat{p}^*_{n_k} \in \{0,1\}) + 2\mathrm{I}(r^*_{n_k} = n_k) + \frac{1}{\sqrt{\widehat{p}^*_{n_k}(1-\widehat{p}^*_{n_k})}} \frac{1}{\sqrt{n_k - r_{n_k}}} \mathrm{I}(0 < \widehat{p}^*_{n_k} < 1, r_{n_k} < n_k),$$

for some constant if $n_k - r^*_{n_k} > 0$. Then $\left|R^*_{n_k}\right| \xrightarrow{p^*} 0$ follows because $1/\sqrt{n_k - r^*_{n_k}} \xrightarrow{p^*} 0$ and $\widehat{p}^*_{n_k} \xrightarrow{p*} p_0 \in (0,1)$ as $n_k \to \infty$. It also holds that

$$\frac{\sqrt{n_k - r^*_{n_k}}}{\sqrt{n_k - r_{n_k}}} \frac{\sqrt{n_k - r_{n_k}}(\widehat{p}^*_{n_k} - \widehat{p}_{n_k})}{\sqrt{\widehat{p}_{n_k}(1-\widehat{p}_{n_k})}} \xrightarrow{d^*} \sqrt{v_1} Z_1,$$

as $n_k \to \infty$ by Slutsky's theorem so that, by the continuous mapping theorem, we have

$$\Pr_*\left(\Psi^*_{n_k} \leq y | n_k - r^*_{n_k}, \widehat{p}^*_{n_k}\right) \xrightarrow{d^*} \Phi_{\mathrm{nor}}\left(y + \sqrt{v_1} Z_1\right).$$

Because the conditional probability is bounded by 1, we have that expectations converge in the bootstrap world as

$$\Pr_*(\Psi^*_{n_k} \leq y) = \mathrm{E}_*\left[\Pr_*\left(\Psi^*_{n_k} \leq y | n_k - r^*_{n_k}, \widehat{p}^*_{n_k}\right)\right] \to \mathrm{E}\Phi_{\mathrm{nor}}(y + \sqrt{v_1} Z_1) = \Pr(Z_0 + \sqrt{v_1} Z_1 \leq y)$$

as $n_k \to \infty$. Because the real $y \in \mathbb{R}$ was arbitrary, we have $\Psi^*_{n_k} \xrightarrow{d^*} Z_0 + \sqrt{v_1} Z_1$ as $n_k \to \infty$ (holding with probability 1 along n_k) or

$$\sup_{y \in \mathbb{R}} \left|\Pr_*(\Psi^*_{n_k} \leq y) - \Pr(Z_0 + \sqrt{v_1} Z_1 \leq y)\right| \to 0, \tag{C.3}$$

as $n_k \to \infty$ (with probability 1).

Next we prove that $U^*_n = \mathrm{pbinom}\left(Y^*_n, n - r^*_n, \widehat{p}^*_n\right) \xrightarrow{d^*} \Phi_{\mathrm{nor}}\left(Z_0 + \sqrt{v_1} Z_1\right)$ (in probability). For U^*_n and $u \in (0,1)$, we write

$$\begin{aligned}
\Pr_*\left(U^*_n \leq u\right) &= \Pr_*\left[Y^*_n \leq \mathrm{qbinom}(u, n - r^*_n, \widehat{p}^*_n)\right] \\
&= \Pr_*\left[\Psi^*_n \leq \Phi^{-1}_{\mathrm{nor}}(u) + \frac{L^*_n}{\sqrt{(n - r^*_n)\widehat{p}^*_n(1-\widehat{p}^*_n)}}\right],
\end{aligned}$$

where $L_n^* = \mathrm{qbinom}(u, n-r_n^*, \widehat{p}_n^*) - \Phi_{\mathrm{nor}}^{-1}(u)\sqrt{(n-r_n^*)\widehat{p}_n^*(1-\widehat{p}_n^*)} - (n-r_n^*)\widehat{p}_n^*$. By the Berry-Esseen theorem and $n - r_{n_k}^* \xrightarrow{p^*} \infty$, $\widehat{p}_{n_k}^* \xrightarrow{p^*} p$ along the subsequence n_k (with probability 1), we have

$$\frac{L_{n_k}^*}{\sqrt{(n_k - r_{n_k}^*)\widehat{p}_{n_k}^*(1-\widehat{p}_{n_k}^*)}} = \frac{\mathrm{qbinom}(u, n_k - r_{n_k}^*, \widehat{p}_n^*) - (n_k - r_{n_k}^*)\widehat{p}_{n_k}}{\sqrt{(n_k - r_{n_k}^*)\widehat{p}_{n_k}^*(1-\widehat{p}_{n_k}^*)}} - \Phi_{\mathrm{nor}}^{-1}(u)$$
$$\xrightarrow{p^*} \Phi_{\mathrm{nor}}^{-1}(u) - \Phi_{\mathrm{nor}}^{-1}(u) = 0.$$

By this and (C.3), it follows that

$$\Pr_*\left(U_{n_k}^* \le u\right) \to \Pr\left[Z_0 + \sqrt{v_1}Z_1 \le \Phi_{\mathrm{nor}}^{-1}(u)\right] = \Pr\left[\Phi_{\mathrm{nor}}(Z_0 + \sqrt{v_1}Z_1) \le u\right]$$

as $n_k \to \infty$ for each $u \in (0,1)$ (with probability 1). Because the subsequence $\{n_j\}$ was arbitrary, we have $U_n^* \xrightarrow{d^*} \Phi_{\mathrm{nor}}(Z_0 + \sqrt{v_1}Z_1)$ in probability as $n \to \infty$ or

$$\sup_{u\in(0,1)} \left|\Pr_*(U_n^* \le u) - \Pr[\Phi_{\mathrm{nor}}(Z_0 + \sqrt{v_1}Z_1) \le u]\right| \xrightarrow{p} 0, \tag{C.4}$$

as $n \to \infty$.

The $100(1-\alpha)\%$ upper calibration prediction bound is $\tilde{Y}_{n,1-\alpha}^C$ such that

$$\frac{\tilde{Y}_{n,1-\alpha}^C - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} - \Phi_{\mathrm{nor}}^{-1}(U_{1-\alpha}^*) \xrightarrow{p} 0, \tag{C.5}$$

where $U_{n,1-\alpha}^*$ is the $1-\alpha$ quantile of U_n^*, which follows by the Berry-Esseen theorem applied to the Binomial$(n-r_n, \widehat{p}_n)$ distribution. Let $\kappa_{1-\alpha}$ be the $1-\alpha$ quantile of $\Phi_{\mathrm{nor}}(Z_0 + \sqrt{v_1}Z_1)$. Then $\kappa_{1-\alpha} - U_{n,1-\alpha}^* \xrightarrow{p} 0$ holds by (C.4). Thus from this and (C.5) along with the continuity of Φ_{nor} and Φ_{nor}^{-1}, it follows that

$$\Phi_{\mathrm{nor}}\left[\frac{\tilde{Y}_{n,1-\alpha}^C - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}}\right] \xrightarrow{p} \Phi_{\mathrm{nor}}\left[\Phi_{\mathrm{nor}}^{-1}(\kappa_{1-\alpha})\right] = \kappa_{1-\alpha}. \tag{C.6}$$

Hence, by Lemma 2, (C.6) and Slutsky's theorem, we have

$$\begin{aligned}
&\Pr(Y_n \leq \tilde{Y}^C_{n,1-\alpha}) \\
&= \Pr\left\{ \Phi_{\text{nor}}\left[\frac{Y_n - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \right] - \kappa_{1-\alpha} \leq \Phi_{\text{nor}}\left[\frac{\tilde{Y}^C_{n,1-\alpha} - (n-r_n)\widehat{p}_n}{\sqrt{(n-r_n)\widehat{p}_n(1-\widehat{p}_n)}} \right] - \kappa_{1-\alpha} \right\} \\
&\to \Pr\left[\Phi_{\text{nor}}(Z_0 + \sqrt{v_1}Z_1) - \kappa_{1-\alpha} \leq \kappa_{1-\alpha} - \kappa_{1-\alpha} \right] = 1-\alpha.
\end{aligned}$$

The $1-\alpha$ lower prediction bound $\underset{\sim}{Y}^C_{n,1-\alpha}$ is equal to $\tilde{Y}^C_{n,\alpha}$ or off by 1, so immediately we have $\Pr(Y_n \geq \underset{\sim}{Y}^C_{n,1-\alpha}) = 1-\alpha$. □

Proof of Theorem 3

Proof. To prove Theorem 3 Part 1, let $q_{1-\alpha} \equiv \inf\{z \in \mathbb{R} : \Pr(Z_0 + \sqrt{v_1}Z_1 \leq z) \geq 1-\alpha\}$ denote the $1-\alpha$ quantile of the distribution of $Z_0 + \sqrt{v_1}Z_1$, where Z_1, Z_0 are iid standard normal variables with $v_1 > 0$ as in Lemma 3. Let $Y^*_n \sim$ Binomial$(n - r^*_n, \widehat{p}_n)$ and $\widehat{Q}_{n,1-\alpha} \equiv \inf\{z \in \mathbb{R} : G^*_n(z) \geq 1-\alpha\}$ denote $1-\alpha$ quantile of the bootstrap distribution of $[Y^*_n - (n-r)\widehat{p}_n]/\sqrt{(n-r)\widehat{p}_n(1-\widehat{p}_n)}$ with cdf $G^*_n(z) \equiv \Pr_*\{[Y^*_n - (n-r)\widehat{p}_n]/\sqrt{(n-r)\widehat{p}_n(1-\widehat{p}_n)} \leq z\}$, $z \in \mathbb{R}$. Then, it follows from Lemma 3 Part 2 that $\widehat{Q}_{n,1-\alpha} \xrightarrow{p} q_{1-\alpha}$ as $n \to \infty$. To see this, for any given $\epsilon > 0$, we have $G^*_n(q_{1-\alpha} - \epsilon) \xrightarrow{p} \Pr(Z_0 + \sqrt{v_1}Z_1 \leq q_{1-\alpha} - \epsilon) < 1-\alpha$ and $G^*_n(q_{1-\alpha} + \epsilon) \xrightarrow{p} \Pr\left(Z_0 + \sqrt{v_1}Z_1 \leq q_{1-\alpha} + \epsilon\right) > 1-\alpha$ by Lemma 3 Part 2. Hence, $\Pr\left[G^*_n(q_{1-\alpha} - \epsilon) < 1-\alpha < G^*_n(q_{1-\alpha} + \epsilon)\right] \to 1$ as $n \to \infty$, and this event implies that $q_{1-\alpha} - \epsilon \leq \widehat{Q}_{n,1-\alpha} \leq q_{1-\alpha} + \epsilon$ holds so that $\Pr(|\widehat{Q}_{n,1-\alpha} - q_{1-\alpha}| \leq \epsilon) \to 1$. By definition, the upper prediction bound for Y_n is given by

$$\tilde{Y}^{DB}_{n,1-\alpha} = \widehat{Q}_{n,1-\alpha}\sqrt{(n-r)\widehat{p}_n(1-\widehat{p}_n)} + (n-r)\widehat{p}_n.$$

Then using that $[Y_n - (n-r)\widehat{p}_n]/[\sqrt{(n-r)\widehat{p}_n(1-\widehat{p}_n)}] - \widehat{Q}_{n,1-\alpha} \xrightarrow{d} Z_0 + \sqrt{v_1}Z_1 - q_{1-\alpha}$ (a normal random variable with mean $-q_{1-\alpha}$ and variance $1+v_1$) by Slutsky's theorem from

$\widehat{Q}_{n,1-\alpha} \xrightarrow{p} q_{1-\alpha}$ along with Lemma 3 Part 1, we have that

$$\begin{aligned}
\Pr(Y_n \leq \tilde{Y}^{DB}_{n,1-\alpha}) &= \Pr\left[\frac{Y_n - (n-r)\widehat{p}_n}{\sqrt{(n-r)\widehat{p}_n(1-\widehat{p}_n)}} \leq \widehat{Q}_{n,1-\alpha}\right] \\
&= \Pr\left[\frac{Y_n - (n-r)\widehat{p}_n}{\sqrt{(n-r)\widehat{p}_n(1-\widehat{p}_n)}} - \widehat{Q}_{n,1-\alpha} \leq 0\right] \\
&\to \Pr\left(Z_0 + \sqrt{v_1}Z_1 - q_{1-\alpha} \leq 0\right) \\
&= \Pr\left(Z_0 + \sqrt{v_1}Z_1 \leq q_{1-\alpha}\right) = 1-\alpha.
\end{aligned}$$

This establishes Theorem 3 Part 1. Again, the lower prediction bound $\underset{\sim}{Y}^{DB}_{n,1-\alpha}$ is equal to $\tilde{Y}^{DB}_{n,\alpha}$ or $\tilde{Y}^{DB}_{n,\alpha}$-1, which implies that $\lim_{n\to\infty}\Pr(Y_n \geq \underset{\sim}{Y}^{DB}_{n,1-\alpha}) = 1 - \lim_{n\to\infty}\Pr(Y_n \leq \tilde{Y}^{DB}_{n,\alpha}) = 1-\alpha$.

The proof of Theorem 3 Part 2 follows analogously by replacing $(Y^*_n, \widehat{p}^*_n)$ with $(Y^{**}_n, \widehat{p}^{**}_n)$ and applying Lemma 3 Part 3. □

Section D Extending the Theorems to Multiple-Cohort Data

For multiple-cohort data, we assume that $\lim_{n\to\infty} n_s/n \to c_s \in [0,1]$ exists for $s = 1, ..., S$, where $\sum_{s=1}^{S} c_s = 1$, and then describe some minor modifications needed to the assumptions of Theorem 1 and 2. As in Theorem 1, based on the censored sample, an estimator of $\boldsymbol{\theta} \in \mathbb{R}^q$ is assumed to satisfy $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \text{MVN}(0, \boldsymbol{V}_0)$ as $n \to \infty$, and its bootstrap counterpart approximation is assumed to be distributionally consistent as in Theorem 2. For the multiple-cohort case, the small change to Theorem 1 conditions is that, for each cohort s, we assume that (with respect to the censoring time t^s_c of the cohort) $F(t^s_c; \boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$ with $F(t^s_c; \boldsymbol{\theta}_0) \in (0,1)$, and that the conditional probability $p_s = \pi_s(\boldsymbol{\theta}) = [F(t^s_c + \Delta; \boldsymbol{\theta}) - F(t^s_c; \boldsymbol{\theta})] \,/\, [1 - F(t^s_c; \boldsymbol{\theta})]$ is continuously differentiable in a neighborhood of $\boldsymbol{\theta}_0$ with gradient $\nabla^s_0 = \partial\pi_s(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, where $p_{0,s} = \pi_s(\boldsymbol{\theta}_0) \in (0,1)$; assume also that ∇^s_0 is non-zero for some cohort s with $c_s > 0$. Then, the same statement of Theorem 1 continues to

hold for the multiple-cohort setting upon redefining the term $v_1 > 0$ to be

$$\frac{1}{\sum_{s=1}^{S} c_s[1-F(t_c^s;\boldsymbol{\theta}_0)]p_{0,s}(1-p_{0,s})} \begin{pmatrix} c_1[1-F(t_c^1;\boldsymbol{\theta}_0)] \\ \vdots \\ c_S[1-F(t_c^S;\boldsymbol{\theta}_0)] \end{pmatrix}^t \Gamma_0^t V_0 \Gamma_0 \begin{pmatrix} c_1[1-F(t_c^1;\boldsymbol{\theta}_0)] \\ \vdots \\ c_S[1-F(t_c^S;\boldsymbol{\theta}_0)] \end{pmatrix}$$

where Γ_0 is now a $q \times S$ matrix given by

$$\Gamma_0 \equiv \left[\nabla_0^1 \cdots \nabla_0^S\right].$$

With this change, statements of Lemmas 1-2 also remain valid.

Statements of Theorem 2 and Theorem 3 (and Lemma 3) also continue to hold, if we naturally extend the bootstrap consistency assumption of Theorem 2 so that $\Pr_*(T_1^* \leq t_c^s) \xrightarrow{p} F(t_c^s;\boldsymbol{\theta}_0)$ is assumed for any cohort with $c_s > 0$.

We next describe some (mostly minor) modifications for the proofs to hold under multiple-cohort data.

Extending Lemma 1

Lemma 1 does not change for multiple-cohort data.

Extending Lemma 2

In multiple-cohort data, the ML estimator of conditional probabilities $\widehat{\boldsymbol{p}}_n$ is a vector of length S, where S is the number of cohorts. To extend Lemma 2 Part 1, $\sqrt{n}(\widehat{\boldsymbol{p}}_n - \boldsymbol{p}_0) \xrightarrow{d} X_0$ now holds, where $X_0 \sim \mathrm{MVN}(\mathbf{0}, \Gamma_0^t V_0 \Gamma_0)$, with Γ_0 as above, using the delta method and $\boldsymbol{p}_0 = (p_{0,1}, \ldots, p_{0,S}) \equiv [\pi_1(\boldsymbol{\theta}_0), \ldots, \pi_S(\boldsymbol{\theta}_0)]$. For Lemma 2 Part 2, it holds that

$$\sup_{x \in \mathbb{R}, \|a\|=1} \left|\Pr_*(a^T\sqrt{n}(\widehat{\boldsymbol{p}}_n^* - \boldsymbol{p}_0) \leq z) - \Pr(a^T X_0 \leq z)\right| \xrightarrow{p} 0.$$

The same changes are made to Lemma 2 Part 3.

Extending Lemma 3

The predictand can be written as $Y_n = \sum_{s=1}^{S} Y_s$, where Y_{n_s} is the number of events during future time interval in cohort s. Here n_s is the size of the cohort s. Thus for re-formulating

Lemma 3 Part 1, we write

$$\begin{aligned}A_n \equiv & \frac{Y_n - \sum_{s=1}^{S}(n_s - r_{n_s})\widehat{p}_n^s}{\sqrt{\sum_{s=1}^{S}(n_s - r_{n_s})\widehat{p}_n^s(1-\widehat{p}_n^s)}} \\ = & \frac{Y_n - \sum_{s=1}^{S}(n_s - r_{n_s})p_{0,s}}{\sqrt{\sum_{s=1}^{S}(n_s - r_{n_s})p_{0,s}(1-p_{0,s})}} \quad \frac{\sqrt{\sum_{s=1}^{S}(n_s - r_{n_s})p_{0,s}(1-p_{0,s})}}{\sqrt{\sum_{s=1}^{S}(n_s - r_{n_s})\widehat{p}_n^s(1-\widehat{p}_n^s)}} + \frac{\sum_{s=1}^{S}(n_s - r_{n_s})(p_{0,s} - \widehat{p}_n^s)}{\sqrt{\sum_{s=1}^{S}(n_s - r_{n_s})\widehat{p}_n^s(1-\widehat{p}_n^s)}} \\ \equiv & A_{1n}c_n + A_{2n}\end{aligned}$$

as the sum of two terms, where the second term can be re-written as

$$A_{2n} \equiv \frac{\sum_{s=1}^{S}(n_s - r_{n_s})(p_{0,s} - \widehat{p}_n^s)}{\sqrt{\sum_{s=1}^{S}(n_s - r_{n_s})\widehat{p}_n^s(1-\widehat{p}_n^s)}} = \sum_{s=1}^{S} \frac{\frac{n_s - r_{n_s}}{n}}{\sqrt{\sum_{j=1}^{S}\frac{n_j - r_{n_j}}{n}\widehat{p}_n^j(1-\widehat{p}_n^j)}}\sqrt{n}(p_{0,s} - \widehat{p}_n^s).$$

For each $s = 1, \ldots, S$, note that $(n_s - r_{n_s})/n \to c_s[1 - F(t_c^s; \boldsymbol{\theta}_0)]$ by the weak law of large numbers and that the normal limit of $\sqrt{n}(p_{0,s} - \widehat{p}_n^s)$ is determined by the normal limit of $\sqrt{n}(\widehat{\boldsymbol{p}}_n - \boldsymbol{p}_0)$ through $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$. Additionally, by the consistency of $\widehat{\boldsymbol{\theta}}_n$ and the smoothness of $\pi_s(\boldsymbol{\theta})$ we have that $\widehat{p}_n^s = \pi_s(\widehat{\boldsymbol{\theta}}_n) \xrightarrow{p} \pi_s(\boldsymbol{\theta}_0) = p_{0,s} \in (0,1)$ for each $s = 1, \ldots, S$. Hence, it holds that $A_{2n} \xrightarrow{d} \sqrt{v_1}Z_1$ in distribution (for v_1 given above and standard normal variable Z_1), which follows from Slutsky's theorem and the normal limit of $\sqrt{n}(\widehat{\boldsymbol{p}}_n - \boldsymbol{p}_0)$. The previous arguments also show that

$$c_n \equiv \frac{\sqrt{\sum_{s=1}^{S}(n_s - r_{n_s})p_{0,s}(1-p_{0,s})}}{\sqrt{\sum_{s=1}^{S}(n_s - r_{n_s})\widehat{p}_n^s(1-\widehat{p}_n^s)}} \xrightarrow{p} 1.$$

Finally, as in the original proof of Lemma 3, we may apply the Berry-Esseen theorem to determine a normal limit for the sum A_{1n} appearing in A_n, as a sum of non-identical but independent Bernoulli random variables, conditional on the censored multiple-cohort data. Namely, for fixed real $z \in \mathbb{R}$, it analogously holds that $|\Pr(A_n \leq z) - \mathrm{E}[\Phi_{\text{nor}}(z - A_{2n}/c_n)]| \to 0$ from the Berry-Esseen theorem, where $\mathrm{E}\Phi_{\text{nor}}(z - A_{2n}/c_n) \to \mathrm{E}[\Phi_{\text{nor}}(z - \sqrt{v_1}Z_1)] = \Pr(Z_0 + \sqrt{v_1}Z_1)$ holds for independent standard normal variables Z_0, Z_1 (as $\Phi_{\text{nor}}(z - A_{2n}/c_n)$ converges to $\Phi_{\text{nor}}(z - \sqrt{v_1}Z_1)$ in both distribution and expectation by the continuous mapping theorem combined with $A_{2n}/c_n \xrightarrow{d} \sqrt{v_1}Z_1$).

Lemma 3 Part 2 and Lemma 3 Part 3 remain as re-casts of Lemma 3 Part 1 in the bootstrap world.

Extension of Theorems

In the multiple-cohort case, the predictand has a Poisson-binomial distribution. Then the proof follows the same method as in Binomial case. We only need to replace the standardized form of predictand with the Poisson-binomial counterpart.

Section E Simulation Results of Section 8 & the Bearing Cage Data

Section 8 provides a summary of the results from our simulation study. This section provides a graphical summary of the results at the other factor-level combinations used in the study.

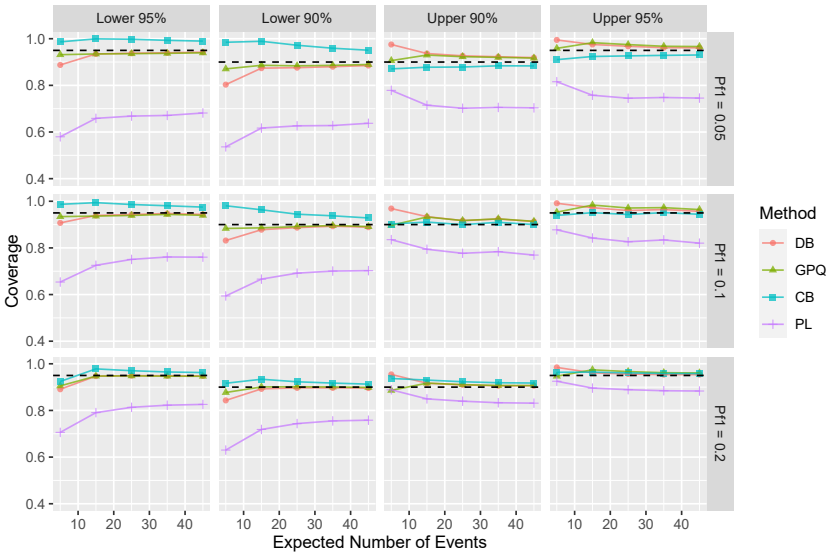

Figure 6: Coverage probabilities versus expected number of events for the direct-bootstrap, GPQ-bootstrap, calibration-bootstrap, and plug-in methods when $d = p_{f2} - p_{f1} = 0.1$ and $\beta = 0.5$.

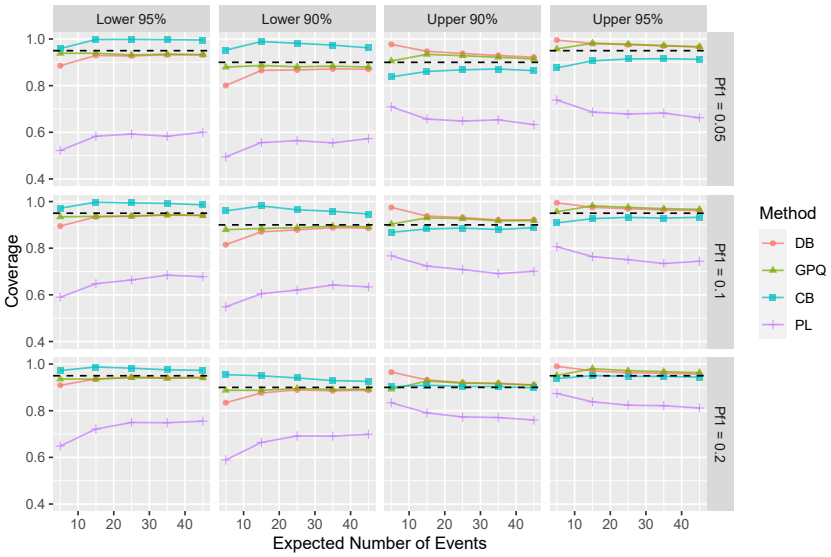

Figure 7: Coverage probabilities versus expected number of events for the direct-bootstrap, GPQ-bootstrap, calibration-bootstrap, and plug-in methods when $d = p_{f2} - p_{f1} = 0.2$ and $\beta = 0.5$.

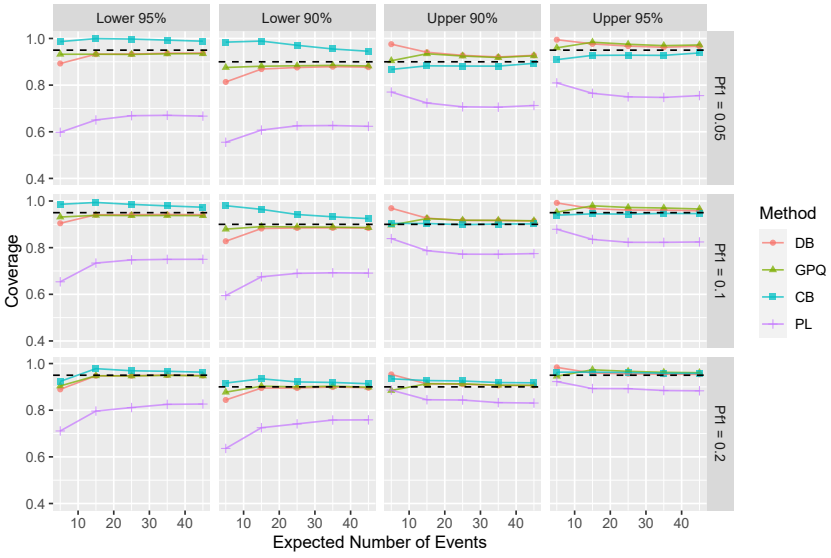

Figure 8: Coverage probabilities versus expected number of events for the direct-bootstrap, GPQ-bootstrap, calibration-bootstrap, and plug-in methods when $d = p_{f2} - p_{f1} = 0.1$ and $\beta = 0.8$.

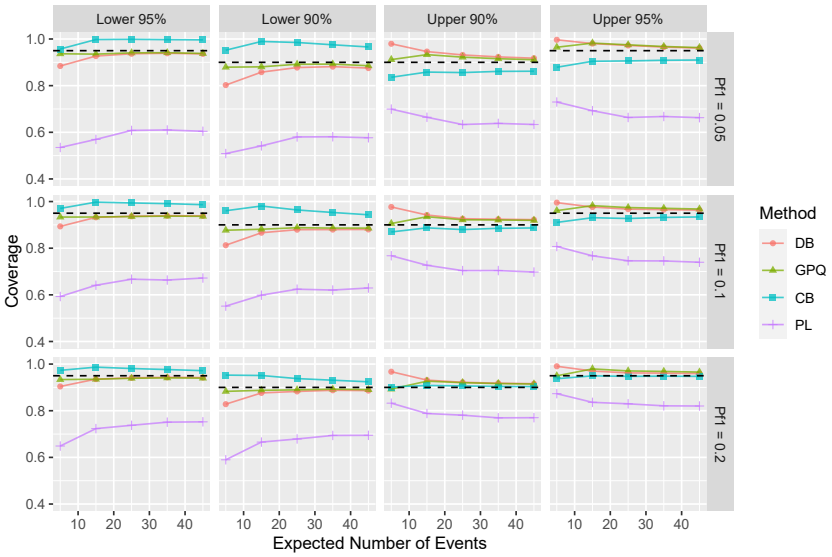

Figure 9: Coverage probabilities versus expected number of events for the direct-bootstrap, GPQ-bootstrap, calibration-bootstrap, and plug-in methods when $d = p_{f2} - p_{f1} = 0.2$ and $\beta = 0.8$.

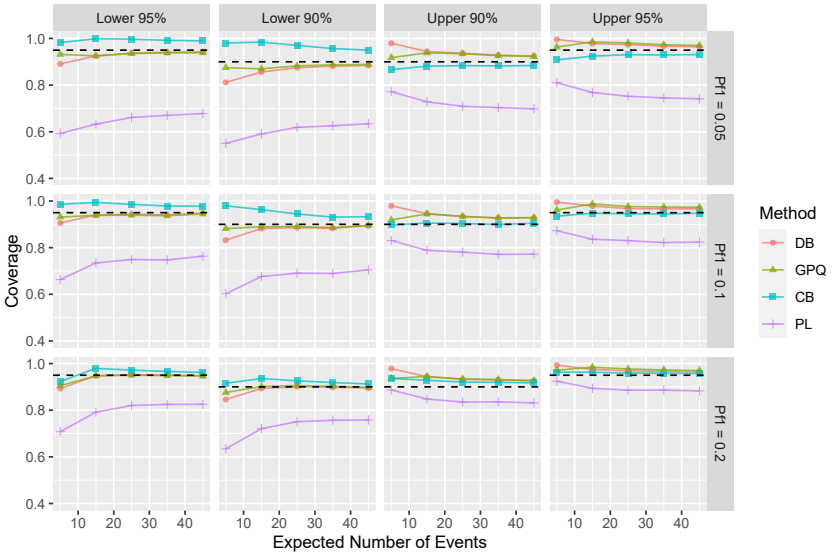

Figure 10: Coverage probabilities versus expected number of events for the direct-bootstrap, GPQ-bootstrap, calibration-bootstrap, and plug-in methods when $d = p_{f2} - p_{f1} = 0.1$ and $\beta = 2$.

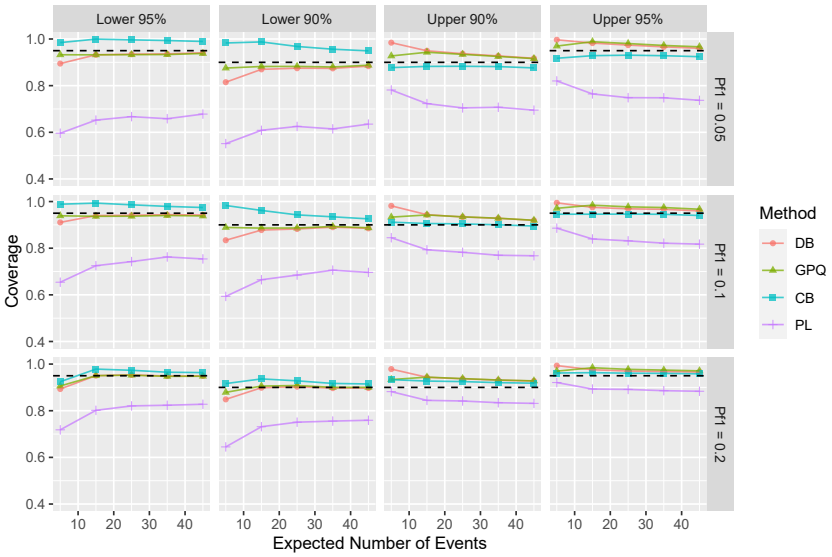

Figure 11: Coverage probabilities versus expected number of events for the direct-bootstrap, GPQ-bootstrap, calibration-bootstrap, and plug-in methods when $d = p_{f2} - p_{f1} = 0.1$ and $\beta = 4$.

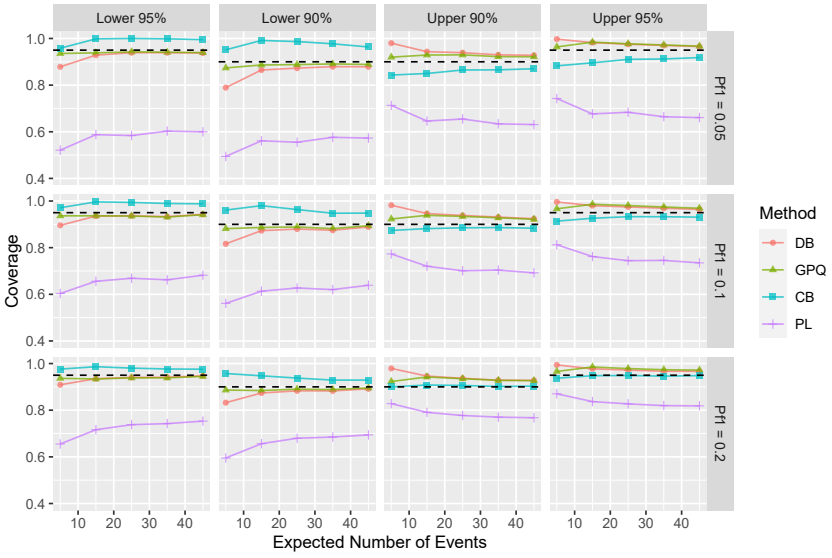

Figure 12: Coverage probabilities versus expected number of events for the direct-bootstrap, GPQ-bootstrap, calibration-bootstrap, and plug-in methods when $d = p_{f2} - p_{f1} = 0.2$ and $\beta = 4$.

Table 6 gives the data for the bearing cage example.

Group i	Hours in Service	Group Size n_i	Failed r_i	At Risk $n_i - r_i$	$\hat{p}_i$	$(n_i - r_i) \times \hat{p}_i$
1	50	288	0	288	0.000763	0.2196
2	150	148	0	148	0.001158	0.1714
3	250	125	1	124	0.001558	0.1932
4	350	112	1	111	0.001962	0.2178
5	450	107	1	106	0.002369	0.2511
6	550	99	0	99	0.002778	0.2750
7	650	110	0	110	0.003189	0.3508
8	750	114	0	114	0.003602	0.4106
9	850	119	0	119	0.004016	0.4779
10	950	128	0	128	0.004432	0.5673
11	1050	124	2	122	0.004848	0.5915
12	1150	93	0	93	0.005266	0.4898
13	1250	47	0	47	0.005685	0.2672
14	1350	41	0	41	0.006105	0.2503
15	1450	27	0	27	0.006525	0.1762
16	1550	12	1	11	0.006946	0.0764
17	1650	6	0	6	0.007368	0.0442
18	1750	0	0	0	0.007791	0
19	1850	1	0	1	0.008214	0.0082
20	1950	0	0	0	0.008638	0
21	2050	2	0	2	0.009062	0.0181
Total	1703		6			5.062

Table 6: Bearing cage data: future-failure risk analysis for the next year (300 hours of service per unit).

Section F Additional Distributional Comparisons

Section 10 of the main manuscript compares the Weibull, lognormal, and Fréchet distributions (as examples of event or failure time models) and presents the plots for $\beta = 2$. Additional plots for $\beta = 1, 4$ are given here. The intent of these plots is to show how different models may vary in their probabilities of future failure events, while having similar percentiles in an initial range where data are observed (e.g., prior to a censoring time t_c for single-cohort data).

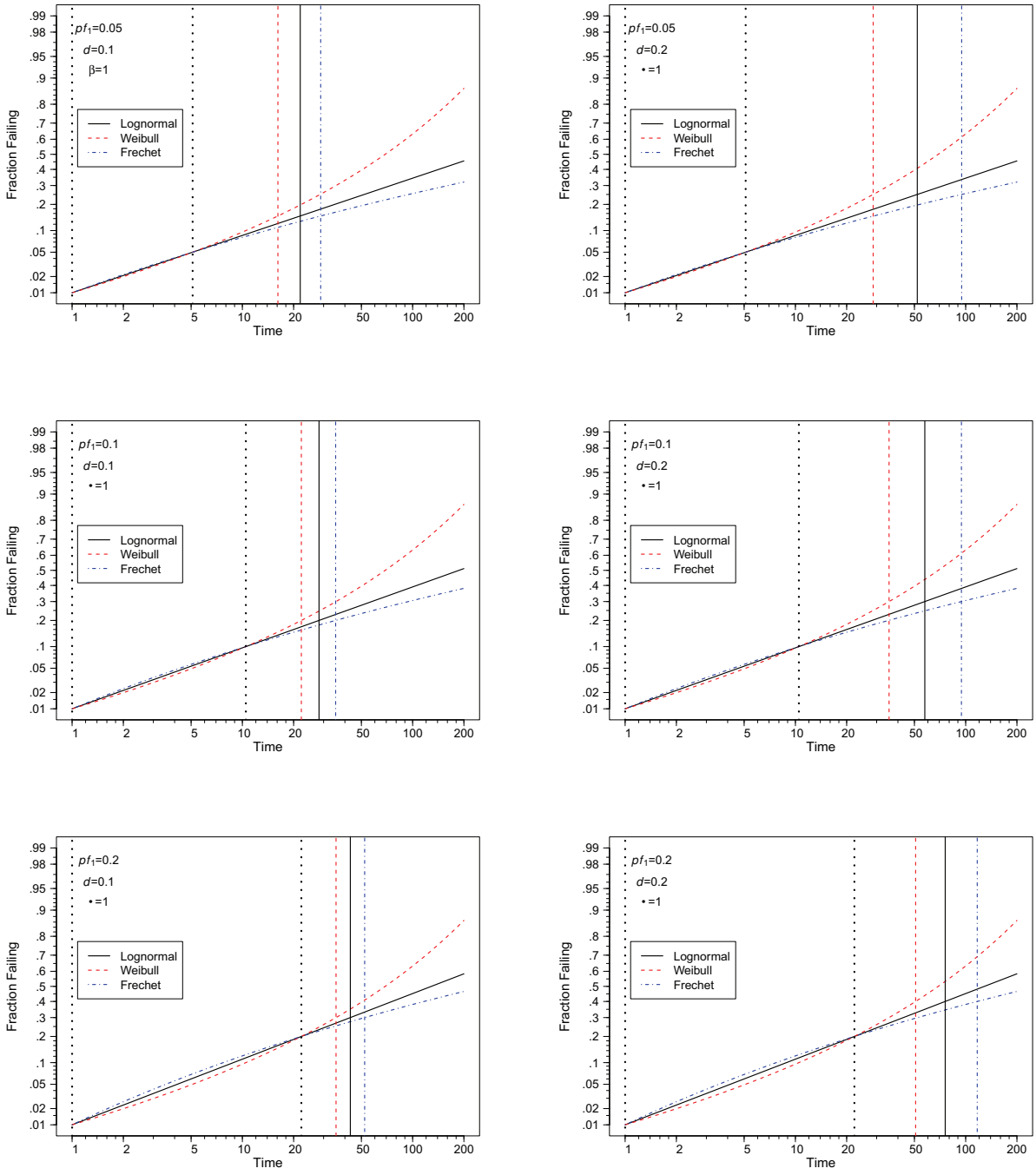

Figure 13: Distributional comparisons for $\beta = 1$. The two vertical dotted lines on the left indicate the points in time where all three distributions have the same 0.01 and p_{f1} quantiles. The three vertical lines on the right indicate the times at $p_{f2} = p_{f1} + d$ for three distributions.

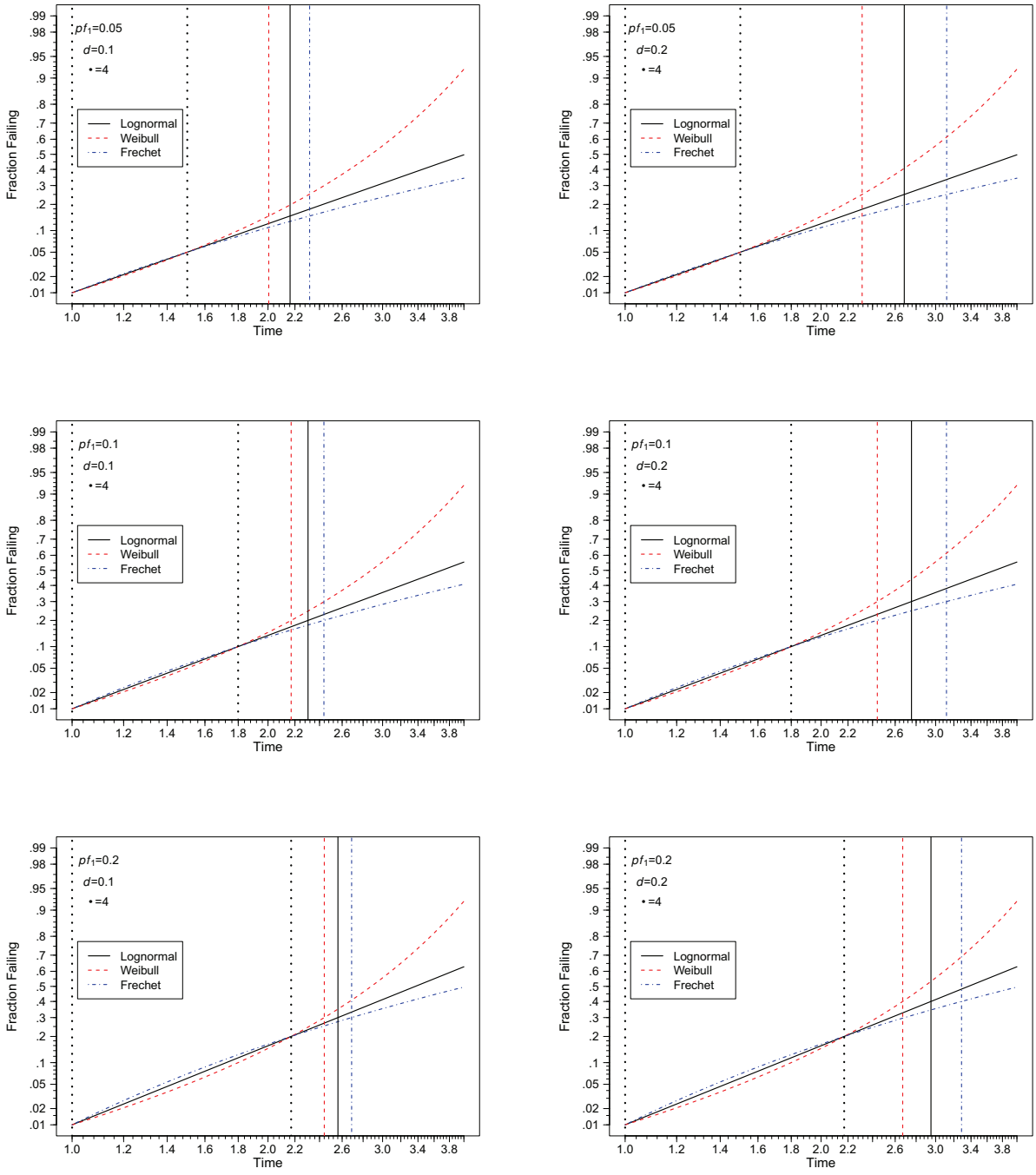

Figure 14: Distributional comparisons for $\beta = 4$. The two vertical dotted lines on the left indicate the points in time where all three distributions have the same 0.01 and p_{f1} quantiles. The three vertical lines on the right indicate the times at $p_{f2} = p_{f1} + d$ for three distributions.