# Predicting the Number of Future Events

Qinglong Tian

Joint work with Fanqi Meng, Daniel Nordman, William Meeker
Department of Statistics, Iowa State University

August 15, 2020

# Predicting the Number of Future Events

**Overview**

- We identify a particular type of prediction problem called within-sample prediction where the future random variable is related to the same sample that provided the original (censored) data.
- We show that the naive plug-in method is not asymptotically correct for within-sample prediction.
- We provide solutions for constructing prediction intervals in within sample prediction and prove that they are asymptotically correct.

# Within Sample Prediction

### Examples

$n = 10{,}000$ units of product were put into service and over the next 48 months, 80 failures occurred and the failure times were recorded. Management requested an upper prediction bound on the number of failures among the remaining 9920 units during the next 12 months.

- 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 we wish to predict (i.e., a count) relates to the same sample that provided the original (censored) data.

# Notations

- Let $(T_1, ..., T_n)$ be a sample from a parametric distribution $F(t; \boldsymbol{\theta})$ having support on the positive real line.
- The available data may then be expressed by $D_i = (\delta_i, T_i^{obs}), i = 1, ..., n$, where $\delta_i = \mathrm{I}(T_i \le t_c)$ and $T_i^{obs} = T_i\delta_i + t_c(1 - \delta_i)$.
- The observed number of events (uncensored units) in the sample will be denoted by $r_n = \sum_{i=1}^{n} \mathrm{I}(T_i \le t_c)$. For a future time $t_w > t_c$, let $Y_n = \sum_{i=1}^{n} \mathrm{I}(T_i \in (t_c, t_w])$ denote the (future) number of values from $T_1, ..., T_n$, that occur in the interval $(t_c, t_w]$, $\boldsymbol{\theta} \in \mathbb{R}^q$.
- The goal is to construct a prediction interval for $Y_n$ based on the observed data $\boldsymbol{D}_n = (D_1, ..., D_n)$ when $\boldsymbol{\theta}$ is unknown.

# The Plug-in Method

The conditional distribution of $Y_n$ is then $\mathrm{Binomial}(n - r_n, p)$ given the observed data $\boldsymbol{D}_n = (D_1, ..., D_n)$, 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

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

Let $\widehat{\boldsymbol{\theta}}_n$ denote an estimator of $\boldsymbol{\theta}$ based on $D_n$, then a plug-in estimator $\widehat{p}_n = \pi(\widehat{\boldsymbol{\theta}}_n)$ of the conditional probability $p$ follows from (1). The plug-in upper prediction bound is

$$\tilde{Y}^{PL}_{n,1-\alpha} = \inf\{y \in \{0\} \cup \mathbb{Z}^+; \mathrm{pbinom}(y, n - r_n, \widehat{p}_n) \geq 1 - \alpha\}.$$

# Regular Prediction Problem

However, plug-in estimation of prediction distributions has only been considered (& shown to be valid) for regular prediction problems by our following definition:

## Definition

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., $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$) & $G_n(y|D_n; \theta)$ is the conditional cdf of $Y|D_n$.

# Failure of the Plug-in Method

**Theorem 1**: The within sample prediction is **not regular** and the plug-in method is **not** asymptotically correct:

1. $\sup_{y\in\mathbb{R}}\left|G_n(y|\boldsymbol{D}_n,\boldsymbol{\theta}_0)-G_n(y|\boldsymbol{D}_n,\widehat{\boldsymbol{\theta}}_n)\right|\xrightarrow{d}$ $1-2\Phi_{\mathrm{nor}}(\sqrt{v_1}|Z_1|/2)$, for $Z_1\sim N(0,1)$ and $v_1$ is a function of $t_c$ and $\boldsymbol{\theta}_0$.
2. The plug-in upper prediction bound $\tilde{Y}^{PL}_{n,1-\alpha}$ generally fails to have an asymptotically correct coverage:

$$\lim_{n\to\infty}\Pr(Y_n\leq\tilde{Y}^{PL}_{n,1-\alpha})=\Lambda_{1-\alpha}(v_1)\in(0,1)\quad\text{such that}$$

$$\operatorname{sgn}\left[\Lambda_{1-\alpha}(v_1)-(1-\alpha)\right]=\begin{cases}1 & \text{if } \alpha\in(1/2,1)\\ 0 & \text{if } \alpha=1/2\\ -1 & \text{if } \alpha\in(0,1/2),\end{cases}$$

# Bootstrap Calibration for Within Sample Prediction

To implement bootstrap calibration, bootstrap method is used to approximate the distribution of $U = \mathbf{pbinom}(Y_n, n - r_n, \widehat{p}_n)$.
For the $100(1-\alpha)\%$ upper prediction bound, the calibrated confidence level is

$$1-\alpha_c = \inf\{u \in [0,1] : \Pr_* \left[ \mathrm{pbinom}(Y_n^{\dagger}, n - r_n^*, \widehat{p}_n^*) \leq u \right] \geq 1-\alpha\},$$

so that the calibrated $100(1-\alpha)\%$ upper prediction bound is $\tilde{Y}^{C}_{n,1-\alpha} = \tilde{Y}^{PL}_{n,1-\alpha_c}$.
Proof in the literature does not apply to within-sample prediction.
We have established that the bootstrap calibration method is asymptotically correct.

# A Different Viewpoint: Predictive Distribution

**Definition**: We use $\widetilde{y}_{1-\alpha}(x_n)$ as a generic notation to denote a $1-\alpha$ upper prediction bound using some prediction method based on data $X_n = x_n$. Then the corresponding predictive distribution $\widetilde{G}_p(\cdot|x_n)$ satisfies

$$\widetilde{G}_p[\widetilde{y}_{1-\alpha}(x_n)|x_n] = 1-\alpha \text{ for } \alpha \in (0,1).$$

- The predictive distribution defines the $1-\alpha$ upper prediction bound as the $1-\alpha$ quantile of the predictive distribution $\widetilde{G}_p(y|x_n)$.
- Correspondingly, given a predictive distribution $\widetilde{G}_p(y|x_n)$, we can treat the $1-\alpha$ quantile of $\widetilde{G}_p(y|x_n)$ as $1-\alpha$ upper prediction bound.

# Alternative: Direct/GPQ Bootstrap Distribution I

- A different type of approach is to construct prediction intervals by constructing the predictive distribution using integration operation. The following two methods are proven to be asymptotically correct.
- Direct Bootstrap: Letting $\Pr_*$ denote bootstrap probability (probability induced by a bootstrap sample $\boldsymbol{D}_n^*$), the direct bootstrap predictive distribution is

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

# Alternative: Direct/GPQ Bootstrap Distribution II

- For log-location-scale distribution, reparameterize the distribution to obtain the location and scale parameters. For example, the Weibull distribution with shape parameter $\beta$ and scale parameter $\eta$, after taking logarithm,the location parameter is $1/\beta$ and the scale parameter is $\exp(\eta)$.
- Letting $\widehat{\boldsymbol{\theta}}_n^* = (\widehat{\mu}_n^*, \widehat{\sigma}_n^*)$ denote a bootstrap version of $\widehat{\boldsymbol{\theta}}_n = (\widehat{\mu}_n, \widehat{\sigma}_n)$, the GPQ bootstrap distribution is the resampling distribution of $\widehat{\boldsymbol{\theta}}_n^{**} = (\widehat{\mu}_n^{**}, \widehat{\sigma}_n^{**})$, where

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

  Similarly, $\widehat{p}_n^{**} = \pi(\widehat{\mu}_n^{**}, \widehat{\sigma}_n^{**})$ can be used to compute the GPQ bootstrap predictive distribution.

# Simulation Study

- Proportion of Failure: $p_{f1} = F(t_c; \boldsymbol{\theta})$.
- Expected Number of Failures: $E(r) = np_{f1}$.
- The Prediction Time Window: $d = p_{f2} - p_{f1} = F(t_w; \boldsymbol{\theta}) - F(t_c; \boldsymbol{\theta})$.
- The Weibull Shape parameter $\beta$ (the scale parameter is set as 1).

# Simulation Study: Weibull Distribution, $d = 0.1$

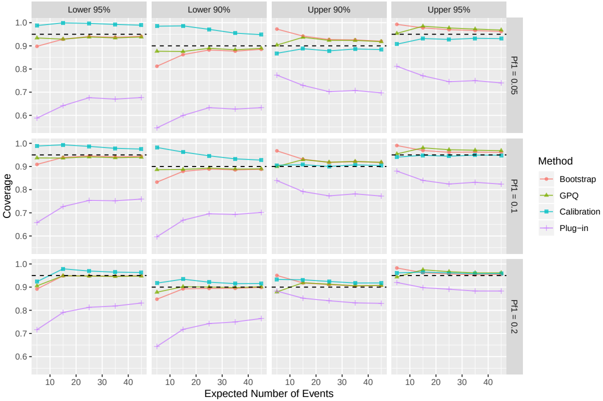

Figure: Coverage probabilities vs. expected number of events for the calibration, bootstrap and approximate GPQ prediction methods when $d = p_{f2} - p_{f1} = 0.1$ and $\beta = 2$.(Plug-in seen as invalid, direct/GPQ work well)

# Multiple Cohort Within-Sample Prediction

| Group | Hours in | Group Size | Failed | At Risk | | |
|---|---|---|---|---|---|---|
| $i$ | Service | $n_i$ | $r_i$ | $n_i - r_i$ | $\hat{p}_i$ | $(n_i - r_i) \times \hat{p}_i$ |
| 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 |

Operationally, the **binom** functions (**pbinom**, **qbinom**, **rbinom**) are replaced by the **poibin** functions (**ppoibin**,**qpoibin**, **rpoibin**).

# Recommendations

- The direct/GPQ bootstrap methods are preferred over bootstrap calibration because of better coverage probability and the bootstrap calibration is computationally unstable.
- Numerical study shows that when number of failures and proportion of failing are small, GPQ bootstrap has better coverage probability than direct bootstrap. The direct bootstrap method tends to be more conservative than the GPQ method on the upper prediction bound but under-coverage compared to the GPQ method on the lower prediction bound.