Chapter 12

Prediction of Future Random Quantities

William Q. Meeker and Luis A. Escobar
Iowa State University and Louisiana State University

Based on the authors' text *Statistical Methods for Reliability Data,* John Wiley & Sons Inc. 1998.

June 21, 2003
14h 53min

Prediction of Future Random Quantities
Chapter 12 Objectives

- Describe problem background and motivation, and some general prediction problem.
- Define probability prediction, naive statistical prediction, and coverage probability.
- Discuss calibrating statistical prediction intervals and pivotal methods.
- Illustrate prediction of the number of future field failures
  - ▶ Single cohort
  - ▶ Multiple cohorts
- Extensions.

Introduction

Motivation: Prediction problems are of interest to consumers, managers, engineers, and scientists.

- A consumer would like to **bound** the failure time of a product to be purchased.
- Managers want to **predict** future warranty costs.
- Engineers want to **predict** the number of failures in a future life test.
- Engineers want to **predict** the number of failures during the following time period (week, month, etc.) of an ongoing life test experiment.

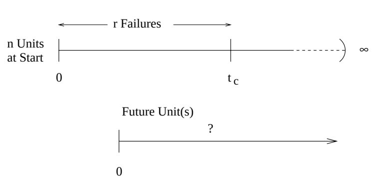
Related Literature

- **Surveys and methods:** Hahn and Nelson (1973), Patel (1989), Hahn and Meeker (1991).
- **Analytical frequentist theory:** Cox (1975), Atwood (1984).
- **Simulation/bootstrap frequentist theory:** Beran (1990), Bai, Bickel, and Olshen (1990), Efron and Tibshirani (1993).
- **Log-location-scale distributions with failure (Type II) censored data—frequentist approach:** Faulkenberry (1973), Lawless (1973), Nelson and Schmee (1979), Engelhardt and Bain (1979), Mee and Kushary (1994).
- **Likelihood theory:** Kalbfleisch (1971).
- **Bayesian theory:** Geisser (1993).

New-Sample Prediction

Based on previous (possibly censored) life test data, one could be interested in:

- **Time** to failure of a **new** item.



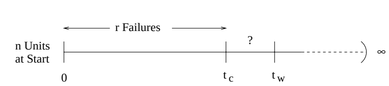
- **Time** until k failures in a **future** sample of m units.
- **Number** of failures by time t_w in a **future** sample of m units.

Within-Sample Prediction

Predict **future** events in a process based on **early** data from the process. Followed n units until t_c and observed r failures. **Data** are first r of n failure times: $t_{(1)} < \ldots < t_{(r)}$.

Want to predict:

- **Number** of additional failures in interval $[t_c, t_w)$.



- **Time** of next failure.
- **Time** until k additional failures.

Needed for Prediction

In general to predict one needs:

- A statistical **model** to describe the population or process of interest. This model usually depends on a set of parameters $\boldsymbol{\theta}$.
- **Information** on the values of the parameters $\boldsymbol{\theta}$. This information could come from
  - ▶ laboratory test.
  - ▶ field data.
- Nonparametric new-sample prediction also possible (e.g., Chapter 5 of Hahn and Meeker 1991).

Probability Prediction Interval
(θ Known)

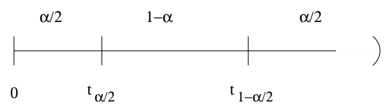
- An exact $100(1-\alpha)\%$ probability prediction interval is (ignoring any data)

$$PI(\alpha) = [\underset{\sim}{T}, \quad \tilde{T}] \;=\; [t_{\alpha/2}, \quad t_{1-\alpha/2}]$$

where $t_p = t_p(\boldsymbol{\theta})$ is the pth quantile of T.

- **Probability of coverage:**

$$\begin{aligned} \Pr[T \in PI(\alpha)] &= \Pr(\underset{\sim}{T} \le T \le \tilde{T}) \\ &= \Pr(t_{\alpha/2} \le T \le t_{1-\alpha/2}) \\ &= 1-\alpha. \end{aligned}$$



Example 1: Probability Prediction for Failure Time of a Single Future Unit Based on Known Parameters

- Assume cycles to failure follows a **lognormal** distribution with **known** parameters $\mu = 4.160, \sigma = .5451$
- A 90% probability prediction interval is

$$\begin{aligned} PI(\alpha) &= [\underset{\sim}{T}, \quad \tilde{T}] = [t_{\alpha/2}, \quad t_{1-\alpha/2}] \\ &= [\exp(4.160 - 1.645 \times .5451), \exp(4.160 + 1.645 \times .5451 \\ &= [26.1, \quad 157.1]\,. \end{aligned}$$

- Then $\Pr(\underset{\sim}{T} \le T \le \tilde{T}) = \Pr(26.1 \le T \le 157.1) = .90$.
- With misspecified parameters, coverage probability may not be .90.

Statistical Prediction Interval
(θ Unknown)

Objective: Want to predict the random quantity T based on a **learning** sample information (DATA).

The random DATA leads parameter estimate $\widehat{\boldsymbol{\theta}}$ and prediction interval $PI(\alpha) = [\underset{\sim}{T}, \quad \tilde{T}]$. Thus $[\underset{\sim}{T}, \quad \tilde{T}]$ and T have a joint distribution that depends on a parameter $\boldsymbol{\theta}$.

Probability of coverage: $PI(\alpha)$ is an **exact** $100(1-\alpha)\%$ prediction interval procedure if

$$\Pr[T \in PI(\alpha)] = \Pr(\underset{\sim}{T} \le T \le \tilde{T}) = 1-\alpha.$$

First we consider evaluation, then specification of $PI(\alpha)$.

Coverage Probabilities Concepts

- **Conditional coverage probability for the interval**:
  For fixed DATA (and thus fixed $\widehat{\boldsymbol{\theta}}$ and $[\underset{\sim}{T}, \quad \tilde{T}]$):

$$\begin{aligned} \mathrm{CP}[PI(\alpha) \mid \widehat{\boldsymbol{\theta}}; \boldsymbol{\theta}] &= \Pr(\underset{\sim}{T} \le T \le \tilde{T} \mid \widehat{\boldsymbol{\theta}}; \boldsymbol{\theta}) \\ &= F(\tilde{T}; \boldsymbol{\theta}) - F(\underset{\sim}{T}; \boldsymbol{\theta}) \end{aligned}$$

  Unknown given $[\underset{\sim}{T}, \quad \tilde{T}]$ because $F(t; \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$.

  Random because $[\underset{\sim}{T}, \quad \tilde{T}]$ depends on $\widehat{\boldsymbol{\theta}}$.

- **Unconditional coverage probability for the procedure**:

$$\begin{aligned} \mathrm{CP}[PI(\alpha); \boldsymbol{\theta}] &= \Pr(\underset{\sim}{T} \le T \le \tilde{T}; \boldsymbol{\theta}) \\ &= \mathrm{E}_{\widehat{\boldsymbol{\theta}}}\left\{\mathrm{CP}[PI(\alpha) \mid \widehat{\boldsymbol{\theta}}; \boldsymbol{\theta}]\right\}. \end{aligned}$$

  In general $\mathrm{CP}[PI(\alpha); \boldsymbol{\theta}] \neq 1-\alpha$.

- When $\mathrm{CP}[PI(\alpha); \boldsymbol{\theta}]$ does not depend on $\boldsymbol{\theta}$, $PI(\alpha)$ is an **exact** procedure.

One-Sided and Two-Sided Prediction Intervals

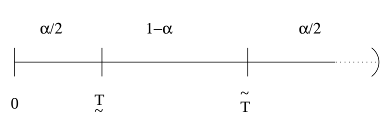
- Combining lower and upper $100(1-\alpha/2)\%$ prediction bounds gives an equal-probability two-sided $100(1-\alpha)\%$ prediction interval.
- If

$$\Pr(\underset{\sim}{T} \le T < \infty) \;=\; 1-\alpha/2 \quad \text{and}$$

$$\Pr(0 < T \le \tilde{T}) \;=\; 1-\alpha/2,$$

  then

$$\Pr(\underset{\sim}{T} \le T \le \tilde{T}) = 1-\alpha.$$



Naive Statistical Prediction Interval

- When $\boldsymbol{\theta}$ is **unknown**, a naive prediction interval is

$$PI(\alpha) = [\underset{\sim}{T}, \quad \widetilde{T}] \;=\; [\widehat{t}_{\alpha/2}, \quad \widehat{t}_{1-\alpha/2}]$$

where $\widehat{t}_p = t_p(\widehat{\boldsymbol{\theta}})$ is the ML estimate of the p quantile of T.

- Coverage probability may be **far** from nominal $1-\alpha$, especially with small samples.

Asymptotic Approximation for CP$[PI(\alpha);\boldsymbol{\theta}]$

As suggested by Cox (1975) and Atwood (1984):

- For the **naive** lower prediction bound:
$PI(\alpha) = [\underset{\sim}{T}, \quad \infty] = [\widehat{t}_\alpha, \quad \infty] = [t_\alpha(\widehat{\boldsymbol{\theta}}), \quad \infty]$, we have

$$\begin{aligned}
\text{CP}\left[PI(\alpha) \mid \widehat{\boldsymbol{\theta}};\boldsymbol{\theta}\right] &= \Pr(\underset{\sim}{T} \le T < \infty;\boldsymbol{\theta}) = g(\alpha,\widehat{\boldsymbol{\theta}};\boldsymbol{\theta}) \\
\text{CP}\left[PI(\alpha);\boldsymbol{\theta}\right] &= \text{E}_{\widehat{\boldsymbol{\theta}}}\left[g(\alpha,\widehat{\boldsymbol{\theta}};\boldsymbol{\theta})\right].
\end{aligned}$$

- Under regularity conditions, using a Taylor expansion of $g(\alpha,\widehat{\boldsymbol{\theta}};\boldsymbol{\theta})$, it follows that

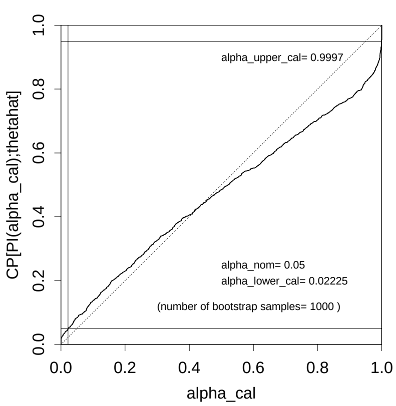
$$\text{CP}[PI(\alpha);\boldsymbol{\theta}] = \alpha + \frac{1}{n}\sum_{i=1}^{k} a_i \left.\frac{\partial g(\alpha,\widehat{\boldsymbol{\theta}};\boldsymbol{\theta})}{\partial \widehat{\theta}_i}\right|_{\boldsymbol{\theta}} + \frac{1}{2n}\sum_{i,j=1}^{k} b_{ij} \left.\frac{\partial^2 g(\alpha,\widehat{\boldsymbol{\theta}};\boldsymbol{\theta})}{\partial \widehat{\theta}_i \partial \widehat{\theta}_j}\right|_{\boldsymbol{\theta}} +$$

where a_i, b_{ij} are elements of vector $\boldsymbol{a}$ and matrix $\boldsymbol{B}$ defined by

$$\begin{aligned}
\text{E}_{\widehat{\boldsymbol{\theta}}}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) &= \boldsymbol{a}(\boldsymbol{\theta}) + o(1/n) \\
\text{E}_{\widehat{\boldsymbol{\theta}}}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\right] &= \boldsymbol{B}(\boldsymbol{\theta}) + o(1/n).
\end{aligned}$$

These are, in general, difficult to compute.

Prediction interval calibration curve lognormal model



Calibrating One-Sided Prediction Bounds

- To **calibrate** the naive one-sided prediction bound, find α_c, such that

$$\begin{aligned}
\text{CP}[PI(\alpha_c);\widehat{\boldsymbol{\theta}}] &= \Pr\left(\underset{\sim}{T} \le T \le \infty;\widehat{\boldsymbol{\theta}}\right) \\
&= \Pr\left[\widehat{t}_{\alpha_c} \le T \le \infty;\widehat{\boldsymbol{\theta}}\right] = 1-\alpha.
\end{aligned}$$

where $\underset{\sim}{T} = \widehat{t}_{\alpha_c}$ is the ML estimator of the t_{α_c} quantile of T.

- Can do this analytically or by simulation.

- When for arbitrary α, CP$[PI(\alpha);\boldsymbol{\theta}]$ does not depend on $\boldsymbol{\theta}$, the **calibrated** $PI(\alpha_c)$ procedure is **exact**.

- For a two-sided interval, do separately for each tail.

Simulation of the Sampling/Prediction Process

To evaluate the coverage probability of $PI(\alpha_0)$ for some specified $0 < \alpha_0 < 1$, do the following:

- Use the assumed model and ML estimates $\widehat{\boldsymbol{\theta}}$ to simulate the sampling **and** prediction process by computing DATA$_j^*$ and T_j^*, $j = 1,\ldots,B$ for a large number B (e.g., $B = 4000$ or $B = 10000$). For each simulated sample/prediction:
- Compute ML **estimates** $\widehat{\boldsymbol{\theta}}_j^*$ from simulated DATA$_j^*$.
- Use α_0 to compute $\underset{\sim}{T}_j^* = \widehat{t}_{\alpha_0}$ from simulated DATA$_j^*$ and compare with the simulated T_j^*. The proportion of the B trials having $T_j^* > \underset{\sim}{T}_j^*$ gives the Monte Carlo evaluation of CP$\left[PI(\alpha_0);\boldsymbol{\theta}\right]$ at $\widehat{\boldsymbol{\theta}}$.
- To obtain a PI with a coverage probability of $100(1-\alpha)$%, find α_c such that CP$[PI(\alpha_c);\widehat{\boldsymbol{\theta}}] = 1-\alpha$.

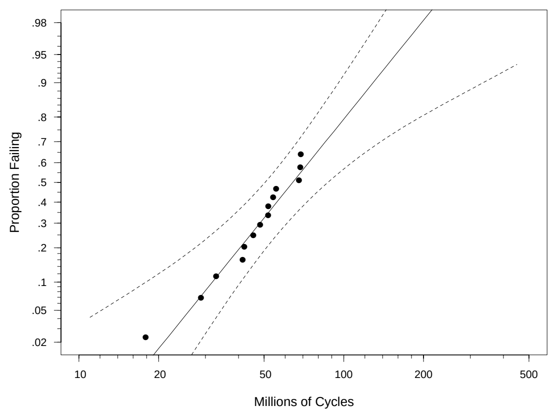
The Effect of Calibration

Result: Beran (1990) showed that, under regularity conditions, with $PI(\alpha_c)$ being a once-calibrated prediction,

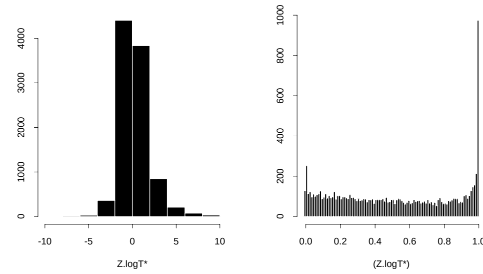
$$\text{CP}[PI(\alpha_c);\boldsymbol{\theta}] = 1-\alpha + O\left(1/n^2\right)$$

and that the order of the approximation can be improved by iterating the calibration procedure.

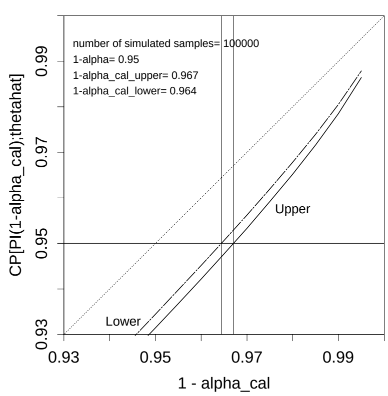
Lognormal probability plot of bearing life test data censored after 80 million cycles with lognormal ML estimates and pointwise 95% confidence intervals



Simulation of the bearing life test censored after 80 million cycles ($n = 23$ and $r = 15$), lognormal model, histograms of pivotal–like $Z_{\log(T^*)} = (\log(T^*) - \widehat{\mu}^*)/\widehat{\sigma}^*$ and $\Phi[Z_{\log(T^*)}]$



Prediction interval calibration function for the bearing life test data censored after 80 million cycles, lognormal model



Example 2: Lower Prediction Bound for a Single Independent Future T Based on Time-Censored (Type I) Data

- Life test run for 80 million cycles; 15 of 23 ball bearings failed. ML estimates of the lognormal parameters are: $\widehat{\mu} = 4.160$, $\widehat{\sigma} = .5451$.
- The naive one-sided **lower** 95% lognormal prediction bound (assuming no sampling error) is:
  $\widehat{t}_{.05} = \exp[4.160 + (-1.645)(.5451)] = 26.1$.
- Need to calibrate to account for sampling variability in the parameter estimates.
- From simulation $\text{CP}[PI(1 - .964); \widehat{\boldsymbol{\theta}}] = .95$
- Thus the **calibrated lower** 95% lognormal prediction bound is

$$\underset{\sim}{T} = \widehat{t}_{.036} = \exp[4.160 + (-1.802)(.5451)] = 24.0$$

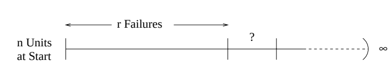
where $z_{.036} = -1.802$.

Comparison of Approximate 90% Prediction Intervals for Bearing Life from a Life Test that was Type I Censored at 80 Million Cycles

	Lognormal	
Naive	[26.1,	157.1]
Calibrated	[24.0,	174.4]

Within-Sample Prediction Predict Number of Failures in Next Time Interval

- The sample DATA are **singly time-censored** (Type I) from $F(t)$. Observe n units until time t_c. Failure times are recorded for the $r > 0$ units that fail in $(0, t_c]$; $n - r$ unfailed at t_c.
- **Prediction problem:** Find an upper bound for the number of future failures, K, in the interval $(t_c, t_w]$, $t_c < t_w$.



Distribution of K and Naive Prediction Bound

- Conditional on DATA, the number of failures K in $(t_c, t_w]$ is distributed as
$$K \sim BIN(n-r, \rho)$$
where
$$\rho = \frac{\Pr(t_c < T \le t_w)}{\Pr(T > t_c)} = \frac{F(t_w;\boldsymbol{\theta}) - F(t_c;\boldsymbol{\theta})}{1 - F(t_c;\boldsymbol{\theta})}.$$

- Obtain $\widehat{\rho}$ by evaluating at $\widehat{\boldsymbol{\theta}}$.

- The naive $100(1-\alpha)\%$ **upper** prediction bound for K is $\widetilde{K}(1-\alpha) = \widehat{K}_{1-\alpha}$, the estimate of the $1-\alpha$ quantile of the distribution of K. This is computed as the smallest integer such that
$$\text{BINCDF}(K, n-r, \widehat{\rho}) > 1-\alpha.$$

Example 3: Prediction of the Number of Future Failures

- $n = 10,000$ units put into service; 80 failures in 48 months. Want an **upper prediction bound** on the number of the remaining
$$n - r = 10000 - 80 = 9920 \text{ units}$$
that will fail between 48 and 60 months.

- Weibull time to failure distribution assumed; ML estimates: $\widehat{\alpha} = 1152,\ \widehat{\beta} = 1.518$ giving
$$\widehat{\rho} = \frac{\widehat{F}(60) - \widehat{F}(48)}{1 - \widehat{F}(48)} = .003233.$$

- Point prediction for the number failing between 48 and 60 months is
$$(n-r) \times \widehat{\rho} = 9920 \times .003233 = 32.07.$$

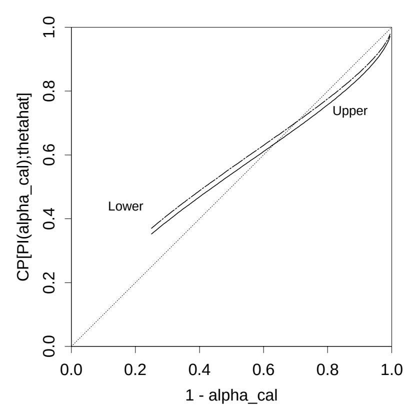
Calibration of the Naive Upper-Prediction Bound for the Number of Field Failures

- Find α_c such that
$$\text{CP}[PI(\alpha_c);\widehat{\boldsymbol{\theta}}] = \Pr\left[K \le \widetilde{K}(1-\alpha_c)\right] = 1-\alpha$$

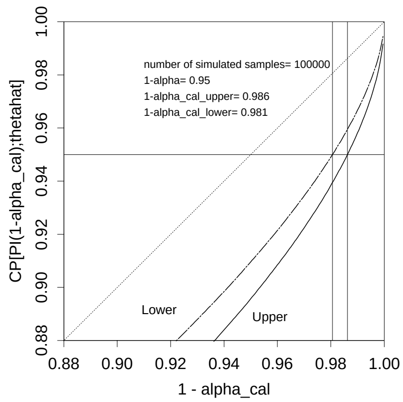
- A Monte Carlo evaluation of the **unconditional** coverage probability is
$$\text{CP}[PI(\alpha_c);\widehat{\boldsymbol{\theta}}] = \frac{1}{B}\sum_{j=1}^{B} P_j$$
where
$$P_j = \text{BINCDF}\left[\underset{\sim}{K}(1-\alpha_c)_j^*; n - r_j^*, \widehat{\rho}\right]$$
is the **conditional** coverage probability for the jth simulated interval evaluated at $\widehat{\rho}$.

- Similar for the lower prediction bound.

Example 3. Calibration functions for upper and lower prediction bounds on the number of future field failures



Example 3. Calibration functions for upper and lower prediction bounds on the number of future field failures

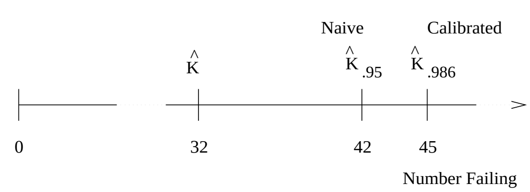


Example 3–Computations

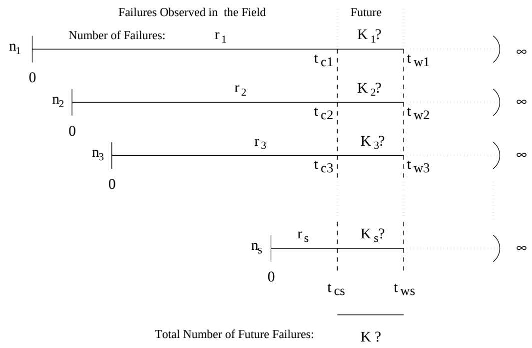
- The **naive** 95% **upper** prediction bound on K is $\widehat{K}_{.95} = 42$, the smallest integer K such that
$$\text{BINCDF}(K, 9920, .003233) > .95.$$

- From simulation $\text{CP}[PI(.9863);\widehat{\boldsymbol{\theta}}] \approx .95$.

- Thus the calibrated 95% **upper** prediction bound on K is $\widetilde{K} = \widehat{K}_{.9863} = 45$, the smallest integer K such that $\text{BINCDF}(K, 9920, .003233) \ge .9863$.



Staggered Entry Prediction Problem



Bearing-Cage Field-Failure Data (from Abernethy et al. 1983)

- A total of 1703 units failed introduced into service over a period of eight years (about 1600 in the past three years).
- Time measured in hours of service.
- Six out of 1703 units failed.
- Unexpected failures early in life mandated a design change.
- How many failures in the next year (point prediction and upper prediction bound requested), assuming 300 hours of service.

Within-Sample Prediction With Staggered Entry

- The objective it to predict **future** events in a process based on several sets of **early** data from the process.
- Units enter the field in **groups** over time. Need to predict the **total** number of **new** failures (in all groups) when unfailed units are observed for an additional period of length Δt.
- For group i, n_i units are followed for a period of length t_{cj} and r_i failures were observed, $i = 1, \ldots, s$.

  DATA$_i$ for set i $(i = 1, \ldots, s)$ are the first r_i of n_i failure times, say $t_{(i1)} < \cdots < t_{(ir_i)}$.

Bearing Cage Data and Future-Failure Risk Analysis

Group i	Hours in Service	n_i	Failed r_i	At Risk $n_i - r_i$	$\widehat{\rho}_i$	$(n_i - r_i) \times \widehat{\rho}_i$
1	50	288	0	288	.000763	.2196
2	150	148	0	148	.001158	.1714
3	250	125	1	124	.001558	.1932
4	350	112	1	111	.001962	.2178
5	450	107	1	106	.002369	.2511
6	550	99	0	99	.002778	.2750
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
17	1650	6	0	6	.007368	.0442
18	1750	0	0	0	.007791	.0000
19	1850	1	0	1	.008214	.0082
20	1950	0	0	0	.008638	.0000
21	2050	2	0	2	.009062	.0181
Total		1703	6			5.057

Distribution of the Number of Future Failures

- Conditional on DATA$_i$, the number of additional failures K_i in group i during interval $(t_{cj}, t_{wi}]$ (where $t_{wi} = t_{cj} + \Delta t$) is distributed as $K_i \sim BIN(n_i - r_i, \rho_i)$ with

$$\rho_i = \frac{\Pr(t_{cj} < T \le t_{wi})}{\Pr(T > t_{cj})} = \frac{F(t_{wi}; \boldsymbol{\theta}) - F(t_{cj}; \boldsymbol{\theta})}{1 - F(t_{cj}; \boldsymbol{\theta})}.$$

- Obtain $\widehat{\rho}_i$ by evaluating $\boldsymbol{\rho} = (\rho_i, \ldots, \rho_s)$ at $\widehat{\boldsymbol{\theta}}$.
- Let $K = \sum_{i=1}^{s} K_i$ be the total number of additional failures over Δt. Conditional on the DATA (and the fixed censoring times) $K \sim$ SBINCDF$(k; \boldsymbol{n} - \boldsymbol{r}, \boldsymbol{\rho})$ a sum of s independent binomials; $\boldsymbol{n} - \boldsymbol{r} = (n_1 - r_1, \ldots, n_s - r_s)$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_s)$.
- A naive $100(1-\alpha)\%$ **upper** prediction bound $\widetilde{K}(1-\alpha)$ is computed as the smallest integer k such that SBINCDF$(k, \boldsymbol{n} - \boldsymbol{r}^*, \widehat{\boldsymbol{\rho}}$ $1 - \alpha$.

Calibration of the Naive Upper Prediction Bound for the Staggered Entry Number of Field Failures

- Find α_c such that

$$\text{CP}[PI(\alpha_c); \widehat{\boldsymbol{\theta}}] = \Pr\left[K \le \widetilde{K}(1-\alpha_c)\right] = 1 - \alpha$$

- A Monte Carlo evaluation of the **unconditional** coverage probability is

$$\text{CP}[PI(\alpha_c); \widehat{\boldsymbol{\theta}}] = \frac{1}{B}\sum_{j=1}^{B} P_j$$

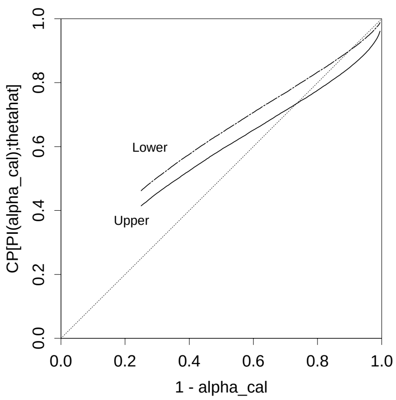
  where

$$P_j = \text{SBINCDF}\left[\underset{\sim}{K}(1-\alpha_c)^*_j; \boldsymbol{n} - \boldsymbol{r}^*, \widehat{\boldsymbol{\rho}}\right]$$

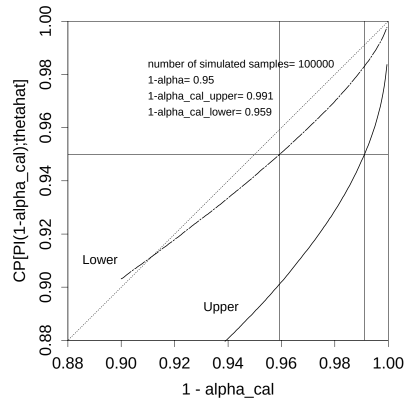
  is the **conditional** coverage probability for the jth simulated interval evaluated at $\widehat{\boldsymbol{\rho}}$.

- Similar for the lower prediction bound.

Example 4: Calibration functions for upper and lower prediction bounds on the number of future field failures with staggered entry



Example 4: Calibration functions for upper and lower prediction bounds on the number of future field failures with staggered entry



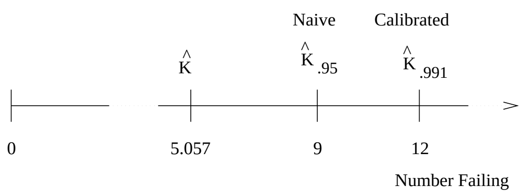
Example 4–Computations

- The **naive** 95% **upper** prediction bound on K is $\widehat{K}_{.95} = 9$, the smallest integer K such that

$$\text{SBINCDF}(K, \boldsymbol{n} - \boldsymbol{r}, \widehat{\boldsymbol{\rho}}) > .95.$$

- From simulation CP$[PI(.9916); \widehat{\boldsymbol{\theta}}] \approx .95$.

- Thus the calibrated 95% **upper** prediction bound on K is $\widetilde{K} = \widehat{K}_{.9916} = 11$, the smallest integer K such that SBINCDF$(K, \boldsymbol{n} - \boldsymbol{r}, \widehat{\boldsymbol{\rho}}) > .9916$.



Concluding Remarks and Future Work

- Methodology can be extended to:
  - Staggered entry with differences among cohort distributions.
  - Staggered entry with differences in remaining warranty period.
  - Modeling of spatial and temporal variability in environmental factors like UV radiation, acid rain, temperature, and humidity.
- Today, the computational price is small; general-purpose software needed.
- Asymptotic theory promises good approximation when not exact; use simulation to verify and compare with other approximate methods.