

**An Imputation Based Approach for Parameter  
Estimation with Application in Industrial Supply Chain Data**

**Samiran Ghosh**

Department of Mathematical Sc., IUPUI, Indianapolis

Email: - [samiran@math.iupui.edu](mailto:samiran@math.iupui.edu)

Special acknowledgement to Dr. Eric Adams,

United Technologies Careers

**Presenting at QPRC, 2009**



## Some Background of the Problem

### Three entities of a Production-Delivery Supply Chain

◆ **Manufacturer**

◆ **Warehouse/Shops**

◆ **Customer**



- To reduce overhead cost often the warehouses/shops are maintained by some third part retailers
- This has immense implication in terms information flow
- Products do fail and warranty is covered by the manufacturer

## Story continued ...

**Note also that,**

- Products are mostly shipped in a batch (not individually)
- When a product fails under warranty, manufacturer knows exactly when the product was sold and for how long does it work
- However only a fraction of the products fail within warranty period
- After a fixed time (say  $T_0$ ) manufacturer does not know for sure what happened to the other products in the same batch?
  1. Are they already sold but working?
  2. Still in the shelf?

**In an Ideal World**

- Manufacturer and retailer has complete transparency
- Manufacturer knows as soon as a product is sold

**Unfortunately this does not happen (often) in reality**

## Story continued ...

### Challenges ahead

- Information is time sensitive and costly and retailers does not yield this individual item specific selling information to the manufacturer
- This poses challenge to the in-house reliability engineers
- Product failure is not only costly but also reliability assessment and future lifetime prediction at an early stage is advantageous
- This has do with customer satisfaction as well as goodwill

### This talk aims to provide solutions related to

- Parameter estimation involving the lifetime of the product
- Optimum usage of the available information
- Computationally efficient approach

Developed methodology is supported by the analysis of Furnace data.

## Some Notations

$N$  :- Number of identical products/units in a batch

$T_0$  :- Fixed observation time generally in terms of year before warranty

$X \sim F_X(\cdot)$  :- Random variable for the installation/selling time

$T \sim F_T(\cdot)$  :- Random variable for the failure time

**We assume**,  $F_X(\cdot)$  and  $F_T(\cdot)$  are completely specified except for the unknown parameters which we need to estimate.

$\Omega = \{i \in \{1, 2, \dots, N\} : X_i + T_i \leq T_0\}$  is the random set which are completely observed before  $T_0$  and  $C = |\Omega|$ .

Essentially  $N - C$  many units are unobserved or censored. However censoring is *ambiguous*.

## Ambiguity in Censoring

As we have noted earlier a unit remains unobserved if

**A. Still in the shelf**

This indicates  $X > T_0$  or Type-1 right censoring on the installation time

**B. It is already sold but still working**

This indicates  $X = x < T_0$  and  $T > T_0 - X$  or Type-1 right censoring on the failure time.

Writing likelihood is rather straightforward if manufacture knows how to distinguish between these two events.

This requires timely information from the retailers about  $X$  (as soon as it is sold). Unfortunately this does not happen for many legacy companies.

## Exact Likelihood with Ambiguous Censoring

The exact likelihood with ambiguous censoring is given by,

$$L(F_X, F_T) = \prod_{i=1}^N [f_{X,T}(x_i, t_i) I\{x_i + t_i \leq T_0\}]^{\tau_i} [P\{X + T > T_0\}]^{1-\tau_i}$$

$$\propto \left\{ \prod_{i \in \Omega} f_{X,T}(x_i, t_i) \right\} \left\{ S_X(T_0) + \int_0^{T_0} S_T(T_0 - x) dF_X(x) \right\}^{N-C}$$

Where  $\tau_i$  an indicator of whether  $i$ -th unit is observed or not.

### Assumptions Made Here

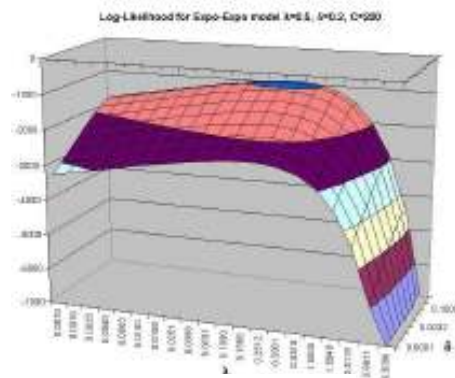
- $X$  and  $T$  have continuous but independent lifetime distribution
- $F_X(\cdot)$  and  $F_T(\cdot)$  are completely specified parametric distribution
- All units in a batch are identical and independent
- There is no significant time lag between *Purchase* and *Installation*
- There is no effect like Idle-Ageing (sitting idle in the warehouse/shop)
- There is no effect due to time of installation (e.g. winter, summer etc.)

## Problem with the Exact Likelihood

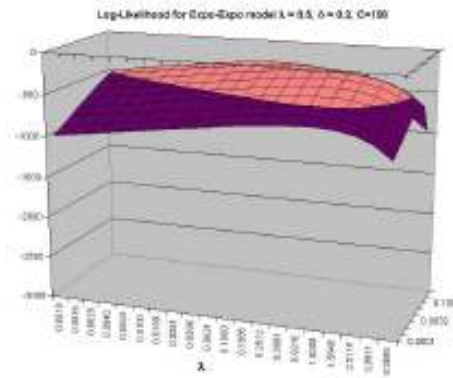
Even if we assume some known distribution for  $X$  and  $T$  (Weibull, Gamma etc.), typically  $c/N$  ratio is 30% or below. With only this much data finding MLE is computationally very challenging.

### A Simple Example,

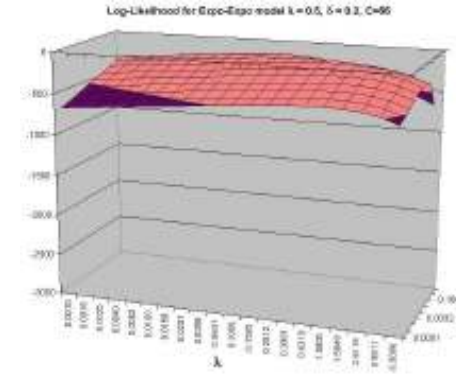
Let  $X \sim \text{Exp}(\lambda = 0.5)$  and  $T \sim \text{Exp}(\delta = 0.2)$  with  $N = 200$



$C=200, T_0 = 12$



$C=100, T_0 = 6$



$C=66, T_0 = 4$

This results in unstable estimate with large variance.

## Standard Practice in the Industry

Reliability engineers (Abertheny, 1996) adopted simplistic approaches

1. Use information pertaining to  $C$  many observed units only.
2. Same as earlier however samples are coming from truncated distributions only.

Both this approaches essentially ignores  $N - C$  many unobserved units completely and as a result only suboptimal and produces over estimation.

### My Work

1. Asymptotic properties of the MLE arising from the exact likelihood
2. A computationally efficient solution to find the MLE via imputation

In this talk I will focus only on the second issue.

## Imputation: What and Why here

To understand the basic intuition lets see the problem again. The information we have about  $C$  many units,

1.  $x|X + T \leq T_0$
2.  $t|X + T \leq T_0$
3.  $x + t|X + T \leq T_0$

With some algebra,

$$\begin{aligned}
 f_X(x|X \leq T_0) &= \frac{f_X(x|X + T \leq T_0)F_{T+X}(T_0)}{F_T(T_0 - x)F_X(T_0)} \\
 &\propto f_X(x|X + T \leq T_0)F_T^{-1}(T_0 - x) \\
 &\propto f_X(x|X + T \leq T_0)\{1 - S_T(T_0 - x)\}^{-1}.
 \end{aligned}$$

**Remark:** Number of samples (if available) from  $\{x|X \leq T_0\}$  will be higher than  $\{x|X + T \leq T_0\}$ . Essentially we have following identity

$$\#\{x|X \leq T_0\} - \#\{x|X + T \leq T_0\} = \#\{x|X \leq T_0 \cap T > T_0 - X\}$$

## Imputation Continued ...

$$\#\{x \mid X \leq T_0\} - \#\{x \mid X + T \leq T_0\} = \#\{x \mid X \leq T_0 \cap T > T_0 - X\}$$

Sample on the right hand side constitute of those units which are installed but still working (hence unobserved due to the reason **B**).

**Idea:** Is there a way to impute those samples (i.e.  $X$  values) so that MLE searching becomes efficient!

**Answer:** We prescribe here a proportional imputation scheme. Note that since  $T_0$  is fixed if we know/impute  $X$ , then  $T = T_0 - X$ .

Note that here is one more problem though; we also do not know out of  $N-C$  many unobserved units for how many we should impute for.

## Proportional Imputation Scheme

Let  $V = \sum_{j=1}^N V_j$  such that,

$$V_j = \begin{cases} 1 & \text{if } j\text{-th unit is installed by } T_0 \\ 0 & \text{otherwise .} \end{cases}$$

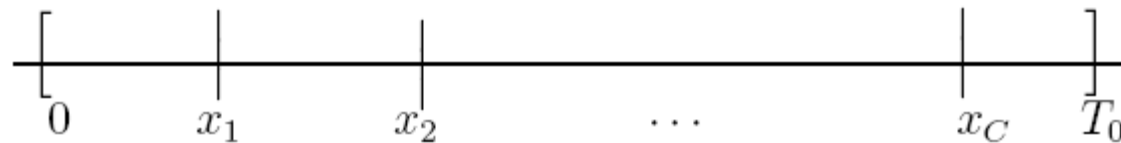
Hence  $V \sim \text{Binomial}(N, F_X(T_0)) \Rightarrow E[V] = NF_X(T_0)$

So we need to impute  $[NF_X(T_0) - C]$  many ( $[\cdot]$  denotes the rounded integer)

But wait, **this makes sense only if we know the parameters in  $F_X(\cdot)$**

For the time being let us assume that we have some crude idea about it.

WLOG,  $C$  many observed units are ordered or  $x_i < x_{i+1}$  for  $i = 1, \dots, C$



These installations produce  $C + 1$  many natural partitions.

## Proportional Scheme Continued ...

Probability of a sample being installed in the interval  $[x_k, x_{k+1}]$  is,

$$P[x_k < X < x_{k+1}] = F_X(x_{k+1}) - F_X(x_k)$$

Probability of such an installed sample remains unobserved,

$$P[T > T_0 - X | x_k < X < x_{k+1}] = \frac{\int_{x_k}^{x_{k+1}} S_T(T_0 - x) f_X(x) dx}{F_X(x_{k+1}) - F_X(x_k)}.$$

Now note the inequality,

$$S_T(T_0 - x_k) \int_{x_k}^{x_{k+1}} f_X(x) dx \leq \int_{x_k}^{x_{k+1}} S_T(T_0 - x) f_X(x) dx \leq S_T(T_0 - x_{k+1}) \int_{x_k}^{x_{k+1}} f_X(x) dx.$$

Using above we would like to approximate,

$$\begin{aligned} I_{k+1} &= P[x_k < X < x_{k+1} \cap T > T_0 - X] \\ &= \frac{S_T(T_0 - x_k) + S_T(T_0 - x_{k+1})}{2} [F_X(x_{k+1}) - F_X(x_k)] \end{aligned}$$

This essentially the joint probability of being installed but not failed.

This approximation works well when intervals are not too large.

Note that,  $I_{k+1}$  gives the probability of an individual item being installed in  $[x_k, x_{k+1}]$ , but remaining unobserved till  $T_0$ .

We have  $C+1$  such intervals (not necessarily equi-spaced)

Hence expected number of unobserved installation in  $[x_k, x_{k+1}]$  is,

$$\alpha_{k+1} = \frac{\{NF_X(T_0) - C\}I_{k+1}}{\sum_{j=0}^C I_{j+1}}$$

with the identity,  $\sum_{k=0}^C \alpha_{k+1} = NF_X(T_0) - C$ . To get the denominator,

Our objective is to get  $\hat{\alpha}_{k+1}$  for  $k=0, \dots, C$  and then use a sampling based approach. We denote the set  $\Gamma = \{i \in \{1, 2, \dots, N\} : X_i \leq T_0 \cap X_i + T_i > T_0\}$  with  $|\Gamma| = \sum_{k=0}^C [\widehat{\alpha_{k+1}}]$  denotes the number of imputed values of  $X$ .

## Likelihood for Installation time

Now combining true observations from  $\Omega$  and imputed ones from  $\Gamma$ , we have  $C+|\Gamma|$  many installations and  $N-C-|\Gamma|$  many type-I right censoring.

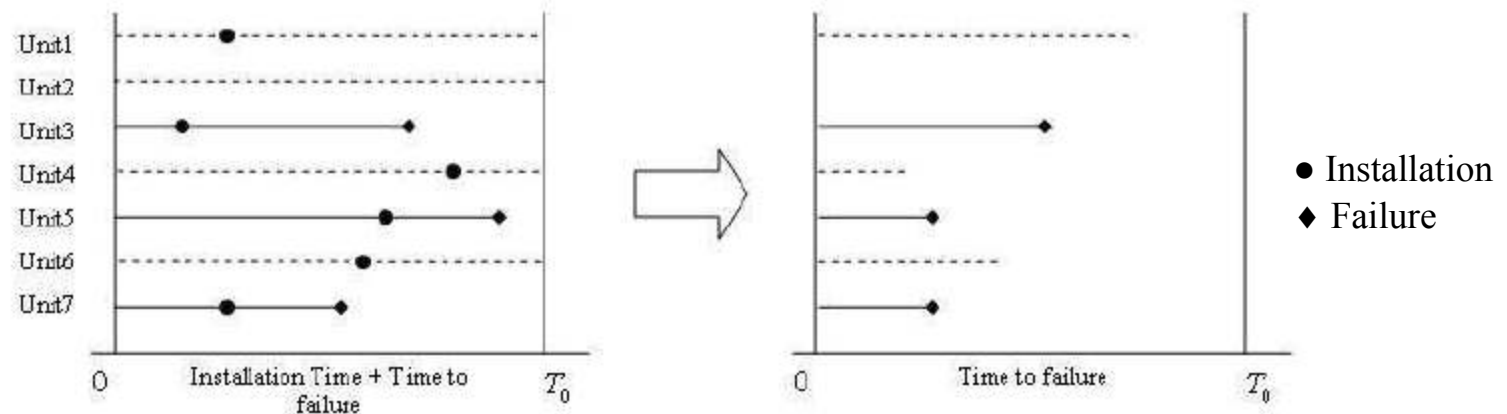
The likelihood is, 
$$L_X = \left\{ \prod_{i \in \Omega \cup \Gamma} f_X(x_i) \right\} S_X(T_0)^{N-C-|\Gamma|},$$

we need to maximize this to obtain the corresponding MLE.

Note: - So far our effort is to characterize the expected number of unobserved installation and distribute them proportionally in different partition of  $[0, T_0]$ .

Next question:- What happen to the failure time ( $T$ ) ?

## Characterization of Failure time



Dashed line unobserved one, solid line observed one

For the time we do not differentiate between imputed  $X$  and true  $X$ .

From the above schematic diagram,

- For unit 1 imputation is done and the remaining is essentially censoring time for  $T$ .
- For unit 2 no imputation is done and hence it is a censoring on  $X$  and does not contribute to  $T$ .
- Obtained failure lifetime is  $T^* = \min\{T, T_0 - X\}$

Let  $\delta$  indicates whether the failure time is censored (0) or not (1) and we have  $n = [NF_X(T_0)]$  many observations (true + imputed).

The likelihood for  $T$  is, 
$$L_T = \prod_{i=1}^n [f_T(t_i^*)]^{\delta_i} [S_T(t_i^*)]^{1-\delta_i}.$$

So far so good but the point is,

- $\alpha_{k+1}$  is unknown unless we know the parameter estimates of  $X$  and  $T$ .

Without that above likelihood has little meaning.

We next proposed an iterative algorithm to get progressively accurate estimate of the  $\alpha_{k+1}$ .

## An Iterative Algorithm

We start with a crude estimate of the parameters in  $X$  and  $T$ . This could be just the MLE of those parameters under the traditional practice that  $C$  many samples came form a truncated distribution.

## The Algorithm

1. Using the current value of distribution parameters find  $\widehat{\alpha}_{k+1}$  for  $k = 0, \dots, C$ . Notably it is quite possible that  $\widehat{\alpha}_{k+1}$  is not an integer. Say  $\widehat{\alpha}_{k+1} = \text{int}(\widehat{\alpha}_{k+1}) + \text{frac}(\widehat{\alpha}_{k+1}) = U_{k+1} + V_{k+1}$ .
2. Draw  $U_{k+1}$  many samples from the interval  $[x_k, x_{k+1}]$  from the distribution  $F_X(\cdot)$  using the current value of the distribution parameters.
3. Draw a sample from a *Bernoulli*( $V_{k+1}$ ) first. If it is 1 draw another sample as in step 2 if 0 skip to next step. Hence the total number of imputed samples are either  $U_{k+1}$  or  $U_{k+1} + 1$ .
4. Re-estimate the parameters of  $X$  using both imputed as well as observed ( $C$  many) samples together via MLE under right censoring .
5. Re-estimate the parameters of  $T$  by using both observed ( $C$  many) and censored samples. Random censoring value for any imputed sample is  $T_0 - X_{\text{imputed}}$ .
6. Return to step 1 until an acceptable convergence is reached on the parameters (or the parameters stabilized).

## Convergence Criteria

Let  $\mu$  is a parameter (either of  $X$  or  $T$ ),  $p$  is a pre-specified integer and  $\varepsilon$  is a pre-specified small value chosen by the end user. We stop the iteration when,

$$\left| \frac{\mu_{i+p} - \mu_i}{\mu_{i+p}} \right| < \varepsilon$$

For the multi-parameter case this need to be achieved for every parameter.

## Some Simulation Studies

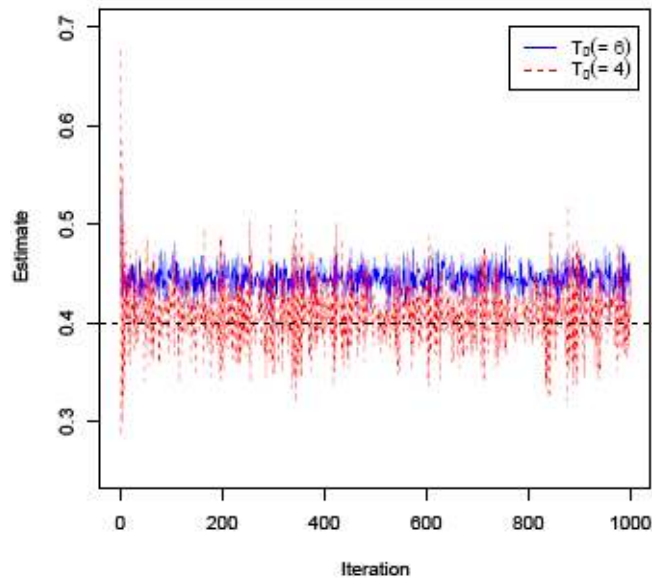
Simulation Setup,

- We choose combinations of Exponential and Weibull distribution
- Instead of stopping the iteration whenever  $\left| \frac{\mu_{i+p} - \mu_i}{\mu_{i+p}} \right| < \varepsilon$ , we ran the iterations 1000 times and throw away the first 100 as non-stabilized value for parameter computation
- We also report the convergence time if we choose  $p = 5$  and  $\varepsilon = 0.0005$

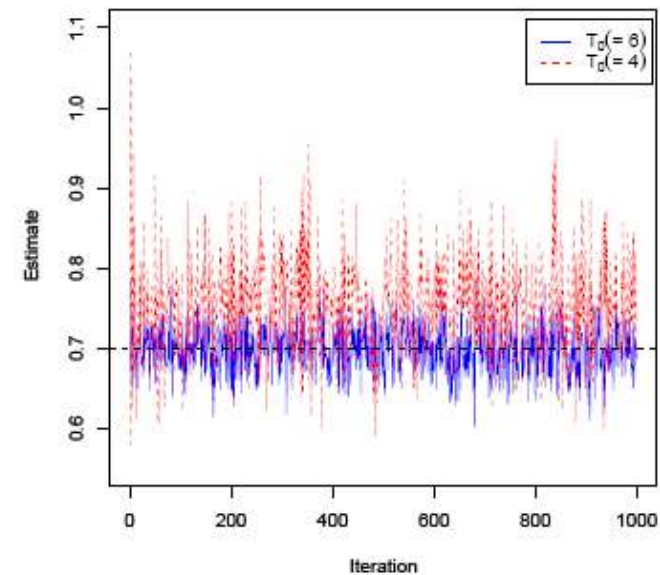
$|D|$  :- Denotes true unobserved installations

Different Distribution	$T_0$	$C$	$ D $	Initial Estimates	Simulation Results	Average # Imputations	Convergence $p = 5, \epsilon = 0.0005$	Time in Second
$X \sim Exp(\lambda = 0.4)$ $T \sim Exp(\delta = 0.7)$	6	170	13	$\lambda_X = 0.53$ $\delta = 0.78$	$\hat{\lambda} = 0.44, \hat{\sigma} = 0.013$ $\hat{\delta} = 0.7, \hat{\sigma}_\delta = 0.03$	18	101	108
$X \sim Exp(\lambda = 0.4)$ $T \sim Exp(\delta = 0.7)$	4	124	43	$\lambda = 0.67$ $\delta = 1.07$	$\hat{\lambda} = 0.41, \hat{\sigma} = 0.03$ $\hat{\delta} = 0.75, \hat{\sigma}_\delta = 0.06$	36	212	155

X estimate plot



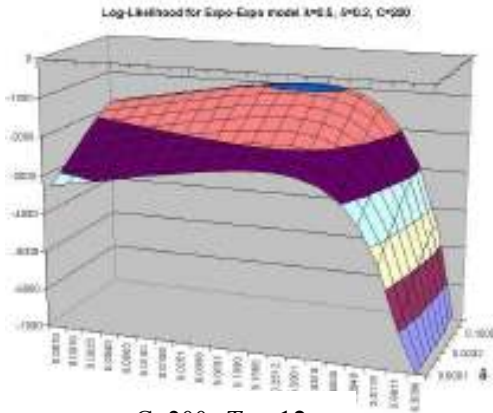
T estimate plot



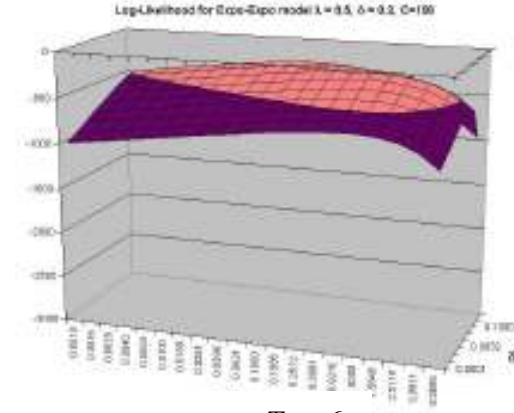
**Note:** - We observe on the average 50% of the all cases.

Let's look back at the likelihood for the case  $X \sim \text{Exp}(\lambda = 0.5)$  &  $T \sim \text{Exp}(\delta = 0.2)$

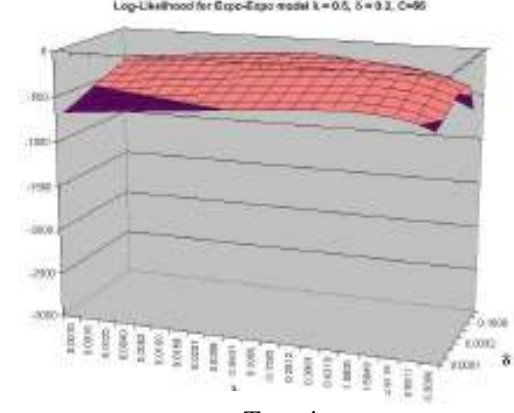
E  
X  
A  
C  
T



$C=200, T_0 = 12$

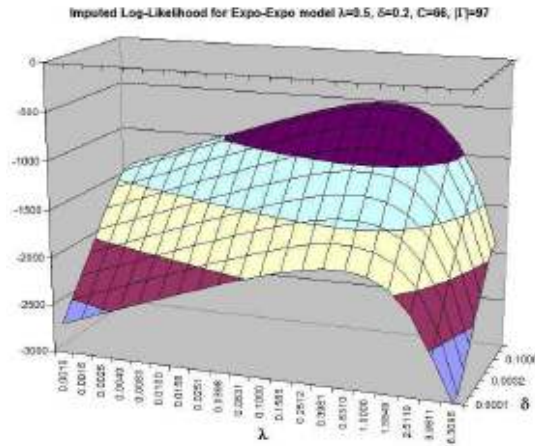


$C=108, T_0 = 6$

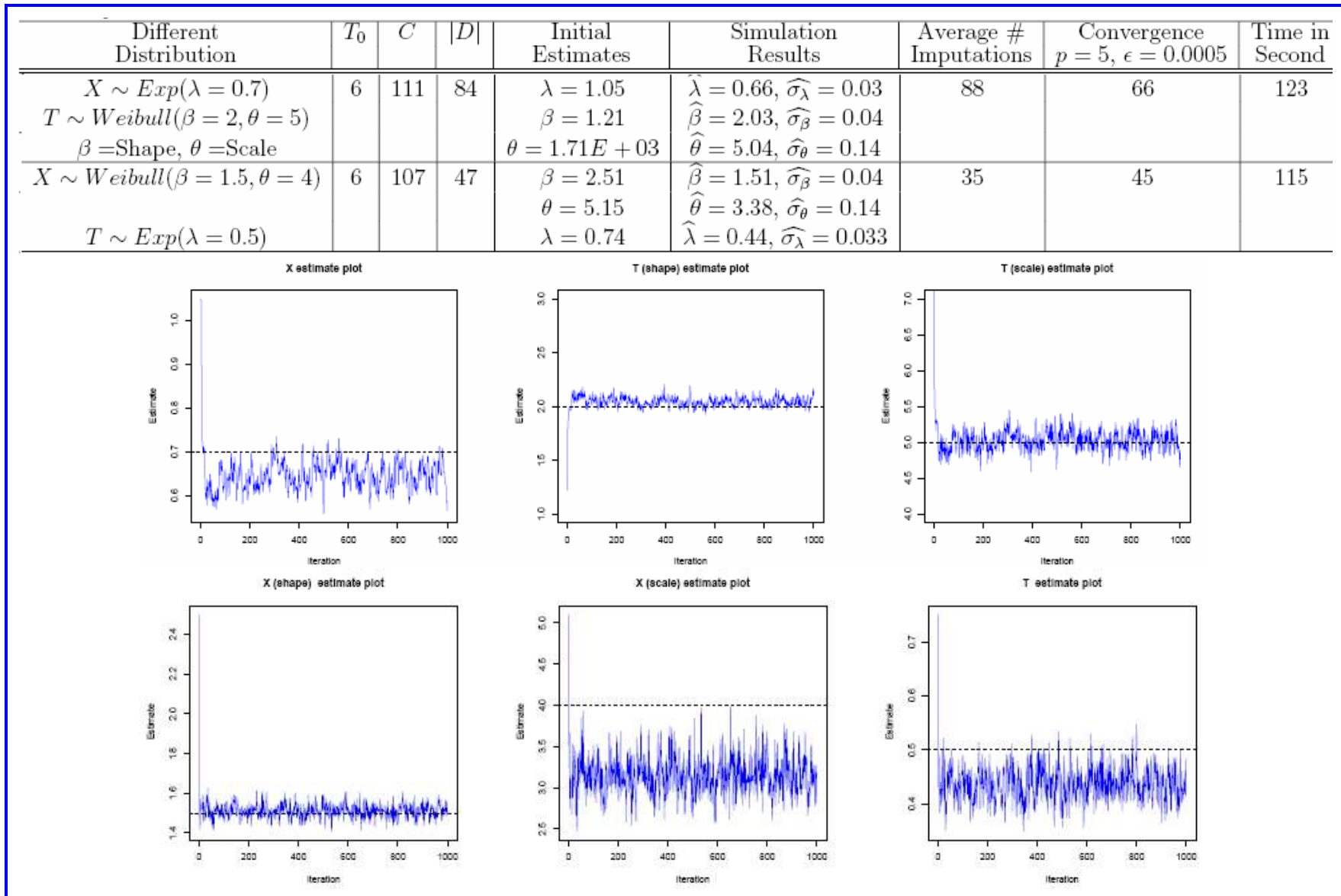


$C=66, T_0 = 4$

I  
M  
P  
U  
T  
E  
D



$C=66, T_0 = 4, |\Gamma| = 97$



## Motivating Example

The data set that we will analyze using the current procedure is coming from an industrial house producing residential furnace components produced during one week in May 2001. We consider single batch with  $N = 400$  units.

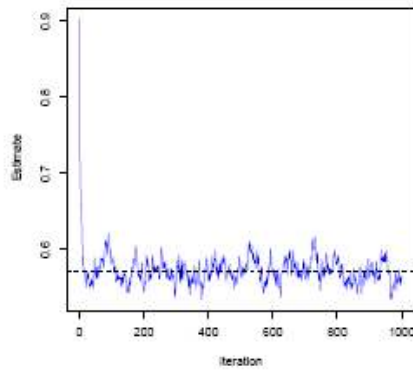
The data consist of  $C = 133$  as observed units which have failed within the observation time of seven years from the manufacturing date.

In the present reliability context the engineers believes that it is appropriate to model installation time ( $X$ ) as exponential while failure time ( $T$ ) as a Weibull distribution (Jager and Bertsche , 2004)

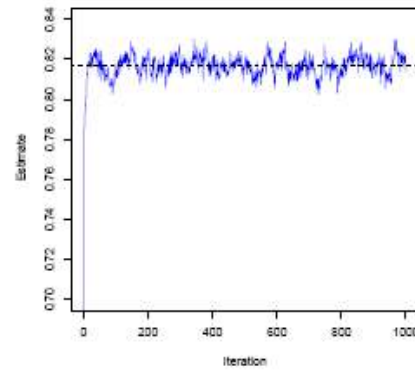
We consider for  $T$  both Weibull and Exponential model. Initial crude estimates are obtained assuming data came from truncated distribution and ignoring all unobserved units.

Distribution	Initial Estimate	Simulation Result	Average # Imputations	Convergence Iteration	Time in Second
$X \sim Exp(\lambda)$	$\lambda = 0.9$	$\hat{\lambda} = 0.57, \hat{\sigma}_{\lambda} = 0.014$	260	167	381
$T \sim Weibull(\beta, \theta)$	$\beta = 0.6, \theta = 3.18$	$\hat{\beta} = 0.81, \hat{\sigma}_{\beta} = 0.004$ $\hat{\theta} = 14.47, \hat{\sigma}_{\theta} = 0.4$			
$T \sim Exp(\delta)$	$\delta = 0.51$	$\hat{\delta} = 0.079, \hat{\sigma}_{\delta} = 0.001$	263	45	421

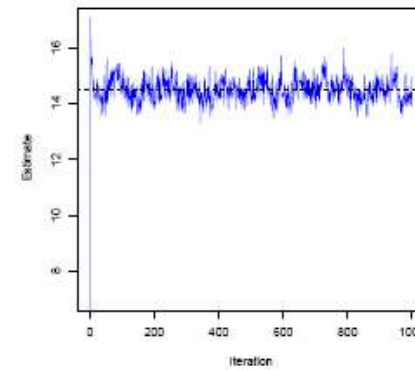
X estimate plot



T (shape) estimate plot

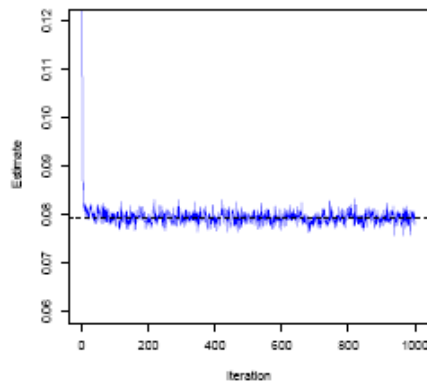


T (scale) estimate plot



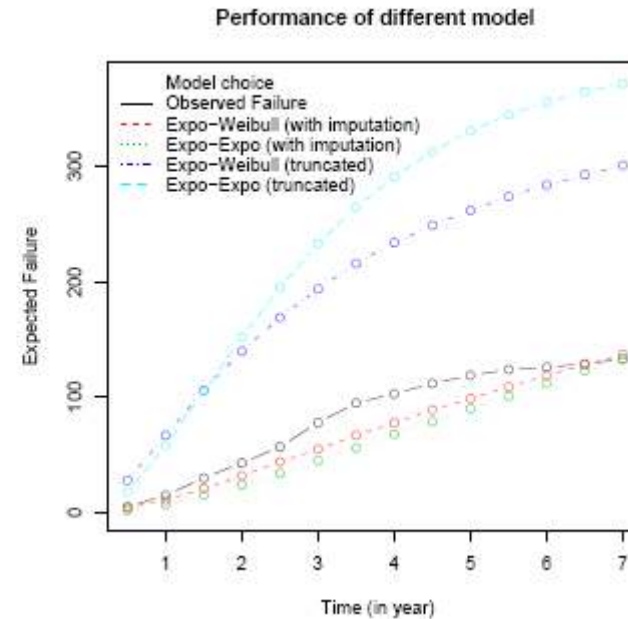
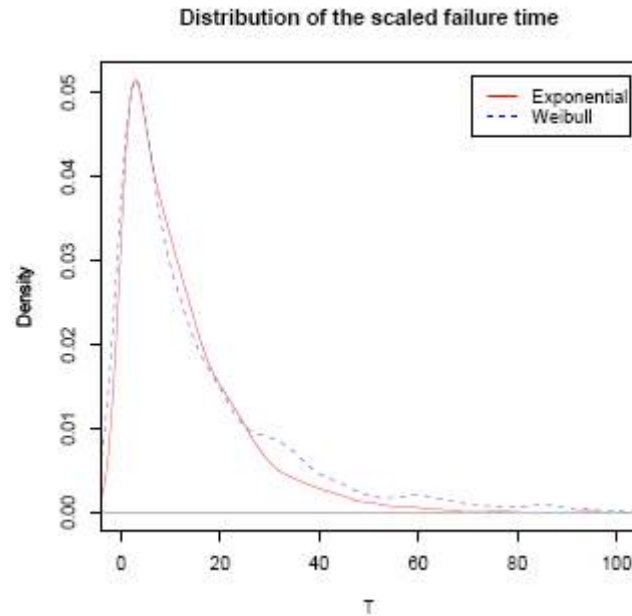
Dashed Line: -  
final reported  
estimate

T estimate plot



Note that Average number of imputation ( $\approx 260$ ) indicates almost all products are sold during that seven year period. This is found to be true latter, though retailers are reluctant to share when they are exactly sold.

For the present case it is found that  $T$  having Weibull or Exponential distributional makes little difference. The density plots of the two distributions of  $T$  are quite similar.



We also compare the predictive performance for all models with observed number of failures by progressively dividing the data for a period of six months. This expected failure number is then compared with the observed failure number.

## Future Work

There exist many directions,

- How to incorporate covariate values (if available)
- How to incorporate failure due to more than one components
- How to analyze more than one batch in a single framework. We need to think of batch effect as well geographical locations where the unit is servicing (e.g. Weather in College Station is not same as Chicago)
- For the most of legacy system in-house engineers have good ideas about the failure time. However what about model misspecification

## Some References : -

1. Abernethy, R.B. (1996). The New Weibull Handbook, 2nd ed. published by Robert B. Abernethy, North Palm Beach, Fl.
2. Damien P. and Walker G. (2001). Sampling truncated normal, beta and gamma distribution. Journal of Computational and Graphical Statistics 10(2), 206-215.
3. Johnson L. G. (1964). The Statistical Treatment of Fatigue Experiments, Amsterdam: Elsevier.
4. Lemon G. (1975). Maximum likelihood estimation for the three parameter Weibull distribution based on censored samples. Technometrics 17(2), 247-254.
5. Johnson L. G. (1964). The Statistical Treatment of Fatigue Experiments, Amsterdam: Elsevier.
6. Meeker W. Q. and Escobar L. A. (1998). Statistical Methods for Reliability Data, New York: John Wiley and Sons.
7. Jager P. and Bertsche B. (2004). A new approach to gathering failure behavior information about mechanical components based on expert knowledge. Reliability and Maintainability Annual Symposium - RAMS, 90-95.
8. Wang W. (2004). Refined rank regression method with censors. Quality and Reliability Engg. Int. 20, 667-678.

**Many Thanks**