

THEORY OF INLIERS: MODELING AND APPLICATIONS

K. Muralidharan

Professor of Statistics and Six Sigma MBB

Department of Statistics,

The Maharajah Sayajirao University of Baroda 390002 India.

WWW.msubaroda.ac.in/academics/~murali

email: muralikustat@gmail.com

University of Bedfordshire, UK, on 07.12.2011

ABOUT ME

Name: Dr. K. Muralidharan

Designation: Professor of statistics and Six sigma MBB

Education: M.Sc. Statistics (1987) Calicut University (Gold medal), India
M.Phil Statistics (1992) Sardar Patel University, India
Ph.D Statistics (1997) Sardar Patel University, India
Post doctoral Fellowship (2003) Academia Sinica, Taipei, Taiwan.

Positions Held: Chair, Department of Statistics, M. S. University of Baroda, India
Chair, Department of Statistics, Bhavnagar University, India (1996-2004)
Dean Faculty of Science, Bhavnagar University (2002-2003)
Coordinator, BCA program, M. S. University of Baroda (since 2009)

Publications: More than 70

Research Guidance (for PhD): 6 (Completed-4, Ongoing-2).

Major Research Projects: 3 (1-DST and 2-UGC).

Industrial Experiences:

Apollo Tyres Baroda

Vardan Foundation (NGO) Baroda

Industrial Jewels Bhavnagar; Acrysil India Ltd Bhavnagar; I&PCL Bhavnagar

Reliance Industries Baroda, Moody ICL, Baroda

Visakhapatnam Steel plant, Andhra Pradesh

SHARP Software Development , Bangalore

Consultancy Experiences:

Dubai electricity and Water Authority, Dubai (Customer Satisfaction Survey & Media Measurement survey)

Department of Economic Development, Dubai (Economic Survey)

Professional Assignments:

Associate Editor of Journal of Probability and Statistical science (JPSS), 2003-2005

Coordinating Editor of JPSS since 2005

Associate Editor of Journal of Investigations in Mathematics (JIM)

Executive Member of Gujarat Statistical Association since 2003

Board Of Studies chairman in Statistics (1996-2004 and 2011)

Visiting Fellow to ISI Delhi and University of Pune

Resource person to many Refresher course and workshop programs

Life member to many Professional bodies.

Research interests:

Statistical inference
Applied Stochastic Process modeling
Life testing and Reliability theory
Probability theory
Industrial statistics
SPC/SQC/Six Sigma/TQM

Awards & recognitions: Honorary appointment to “The Research Board or Advisors” by The American Biographical Institute. Member since 2003, VIJAY RATAN award for excellence in education and Meritorious services-IIFS New Delhi(2005); Best Citizen of India award – International Publishing House, New Delhi (2006), Marquis Who’s Who in Science and Engineering 2006-2007. (Biographical profile included), Commonwealth Academic Fellow (2011)

Conferences/Seminars organized: 8.

Books: 1 (under print with PHI)

PRESENTATION: CONTENTS

- Introduction
- Examples of inliers
- Models
 - (i) Instantaneous failure models
 - (ii) Early failure models
 - (iii) Nearly instantaneous failure models
 - (iv) identified inliers model
 - (v) Labeled slippage inliers model
- Inliers detections
 - (i) Likelihood principle
 - (ii) Information criterions
 - (iii) Bayesian techniques
 - (iv) Sequential procedures
- Testing of hypothesis
- Application

Examples of inliers

1. The distribution of errors in an audit report.
2. Life time experiments on electronic items, where items fails instantaneously and early.
3. Study of tumor characteristics: two variates may be recorded. The first is the absence(0) or presence(1) of a tumor and the second is tumor size measured on a continuous scale. In this problem, it is sometimes of interest to consider a marginal tumor measurement that is 0 with nonzero probability and the other a continuous distribution.

Examples ...

4. Time until remission is of interest in studies of drug effectiveness for treatment of certain diseases. Some patients respond and some do not. The distribution is a mixture of a mass point at 0 and a nontrivial continuous distribution of positive remission times.
5. In different contexts, important problems exist in time-series analysis in which there are mixed spectra containing both discrete and continuous components.
6. The first response time of patients during a medical operation.
7. Number of bugs in a software program

8. Rainfall data:

June: 0.0, 0.0, 0.0, 1.8, 0.0, 0.0, 0.0, 0.0, 0.0, 3.2, 33.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 2.4, 24.2, 33.5, 15.7, 0.0, 0.0, 0.0, 13.8, 0.0, 0.0, 1.0, 0.0

July: 0.0, 44.5, 0.7, 7.0, 14.4, 2.4, 7.3, 4.7, 30.0, 33.3, 12.5, 3.0, 0.0, 2.3, 11.2,
0.0, 6.0, 40.2, 4.9, 0.0, 9.8, 1.3, 9.6, 2.4, 0.6, 0.0, 0.5, 7.2, 1.2, 0.0, 6.0

August: 67.8, 3.2, 1.2, 10.4, 4.3, 1.4, 8.3, 8.1, 0.0, 0.6, 2.1, 3.5, 0.0, 0.8, 4.9, 2.8,
0.0, 0.0, 17.8, 1.3, 0.0, 0.0, 0.0, 0.0, 2.2, 3.8, 0.0, 3.4, 0.0, 0.0, 0.0

September: 1.8, 10.8, 6.0, 8.8, 2.6, 20.1, 2.0, 7.6, 16.4, 2.8, 0.0, 0.0, 15.7, 0.0,
0.0, 20.6, 10.6, 0.0, 0.4, 0.0, 0.0, 1.6, 1.8, 0.0, 6.0, 0.0, 2.5, 0.0, 0.0, 0.0

For more related examples, see Statistical models and analysis in Auditing: Panel on nonstandard mixtures of distributions, Statistical science, vol.4(1), 2-33, 1989.

Introduction

An *inlier* in a set of data to be an observation or subset of observations not necessarily all zeroes, which appears to be inconsistent with the remaining data set.

For example: Consider the following set of observations:

0, 0, 0, 0.12, 0.14, 6.5, 11.9, 14.6, 17.7, 23.9 ...

Here the first 3 observations are called *instantaneous failures*, next two observations may be called *early failures*. Together the first 5 observations may be called *inliers*.

Application areas

- Lifetime experiments
- Reliability Engineering
- Market research
- Clinical trials
- Biological problems

About Outliers

- Originated with the method of least squares.
- Usually after detection they are thrown out.
- Generally appear on the extreme right.
- Associated with mixture distributions.
- Detection can be done using Box-plot
- Problem of Masking effect/swamping effect can influence the conclusions.

Inlier(s) models

Inlier prone models

- Instantaneous failure models
- Early failure models
- Nearly instantaneous failure models
- M_k (identified) inliers models
- L_k (labeled slippage) inliers models
- Inliers as instantaneous and early failures

Instantaneous failure model

$\mathfrak{S} = \{f(x, \theta), x \geq 0, \theta \in \Omega\}$: family of pdf' s.

$$G(x, \theta, p) = \begin{cases} 1 - p, & x = 0 \\ 1 - p + pF(x, \theta), & x > 0 \end{cases}$$

With respect to a measure μ which is the sum of Lebeasgue measure on $(0, \infty)$ and a singular unit measure at the origin

Some references (Instantaneous failure model)

- Aitchison (1955)
- Dahiya and Kleyle (1975)
- Jayade and Prasad (1990)
- Vannman (1991,1995)
- Muralidharan(1999, 2000)
- Muralidharan and Kale (2002)
- Kale (1998)
- Kale and Muralidharan (2006,2008)

Early failure model

$\mathfrak{S} = \{f(x, \theta), x \geq 0, \theta \in \Omega\}$: family of pdf's.

$$G(x, \theta, p) = \begin{cases} 0, & x < \delta \\ 1 - p + pF(\delta, \theta), & x = \delta \\ 1 - p + pF(x, \theta), & x > \delta \end{cases}$$

where δ is known and sufficiently small.

References:

- (i). Kale and Muralidharan (2000)
- (ii). Muralidharan and Lathika (2004, 2005)

Nearly instantaneous failure models

(C. D. Lai, B. C. Khoo, K. Muralidharan and M. Xie, 2007)

- Here the model is written as a complete mixture of two distributions as

$$f(x) = \alpha \Delta_{\delta}(x - x_0) + (1 - \alpha) f_2(x, \theta), 0 < \alpha < 1$$

- Where

$$\Delta_{\delta}(x - x_0) = \begin{cases} \frac{1}{\delta}, & x_0 \leq x \leq x_0 + \delta \\ 0, & \textit{otherwise} \end{cases}$$

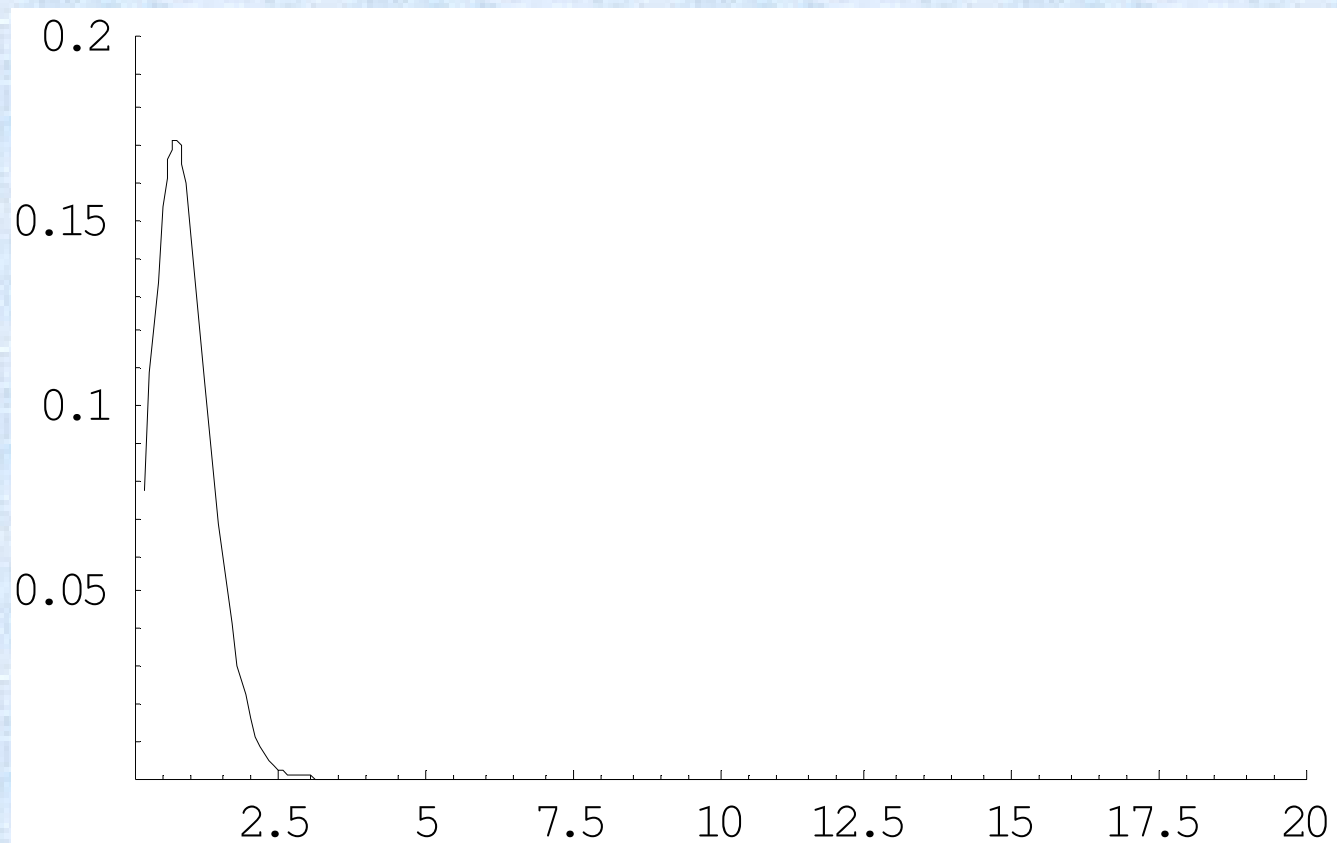
for sufficiently small δ and called the Dirac delta function and $f_2(x, \theta)$ is any other pdf.

The importance of this model are:

- $F(x)$ and hence survival function, $R(x) = 1 - F(x)$ has closed form expression
- $h(x)$, the hazard function also have closed form expression

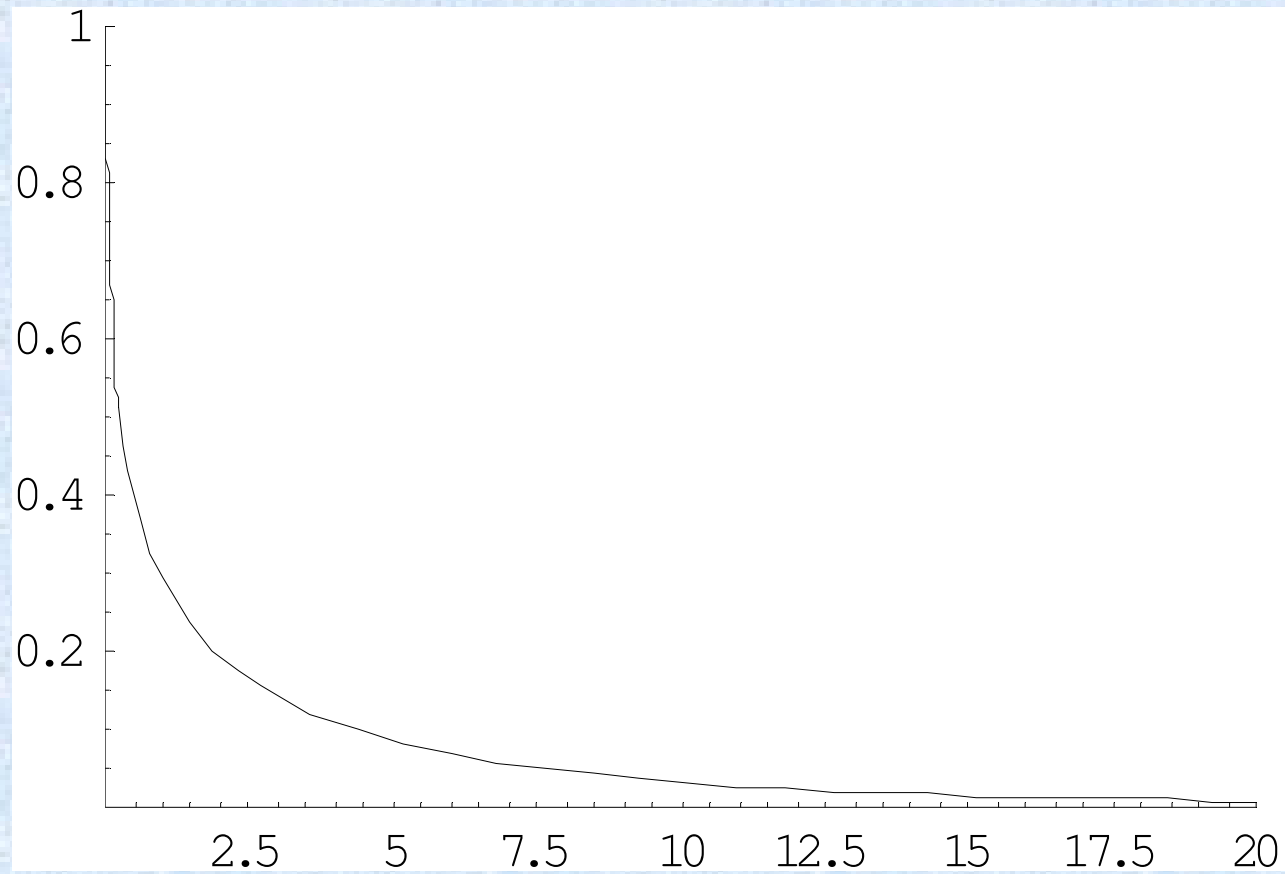
Graph of $f(x)$

$(\theta=1, \beta=2, \delta=0.2, \alpha=0.8)$



Graph of $R(x)$

$(\theta=1, \beta=0.5, \delta=0.2, \alpha=0.2)$



Some assumptions

- Suppose X_1, X_2, \dots, X_n of n observations are such that r of them are independently and identically distributed (i.i.d) from $G \in \mathcal{G}$ and $(n-r)$ of them are i.i.d from $F \in \mathcal{F}$.
- $\frac{\partial G}{\partial F}$ is decreasing in X .
- G is the inlier distribution and F is the target distribution

Likelihood: Identified inliers model

- Here r and the indexing set ν are known.

Therefore,

X_1, X_2, \dots, X_r are iid with $G \in \mathcal{G}$ and

$X_{r+1}, X_{r+2}, \dots, X_n$ are i.i.d. with $F \in \mathcal{F}$.

Hence the likelihood is

$$L(x | g, f, r, \nu) = \prod_{i=1}^r g(x_i) \prod_{i=r+1}^n f(x_i)$$

Likelihood: Labeled slippage inliers model

- Let $X_{(1)} < X_{(2)} < \dots < X_{(r)}$ be the order statistics from G and $X_{(r+1)} < X_{(r+2)} < \dots < X_{(n)}$ be the order statistics from F .

Consider $X_{(1)} < \dots < X_{(r)} < X_{(r+1)} < \dots < X_{(n)}$.Then

$$\begin{aligned}\varphi_r(G, F) &= P[X_{(r)} < X_{(r+1)} \mid G, F) \\ &= \int_{-\infty}^{\infty} (n - r)[G(u)]^r [1 - F(u)]^{n-r-1} dF(u)\end{aligned}$$

Theorem

- Let $X_{(1)} < X_{(2)} < \dots < X_{(n-n_0)}$ be the order statistics and $R_1, R_2, \dots, R_{n-n_0}$ be the corresponding rank order statistics, then

$$\underset{r_1, r_2, \dots, r_k}{\text{Max}} \varphi(r_1, r_2, \dots, r_k) = \varphi(1, 2, \dots, k) \text{ and } (x_{(1)}, x_{(2)}, \dots, x_{(k)})$$

have the maximum probability of being inliers.

(See Kale and Muralidharan, 2006 for proof)

- If $g \sim \exp(\phi)$ and $f \sim \exp(\theta)$, $\phi \ll \theta$, then

$$\begin{aligned}\varphi_r(G, F) &= \int_{-\infty}^{\infty} (n-r) \frac{1}{\theta} [1 - e^{-x/\phi}]^r e^{-(n-r)x/\theta} dx \\ &= \frac{(n-r)\phi}{\theta} B\left(r+1, \frac{(n-r)\phi}{\theta}\right),\end{aligned}$$

where

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$$

Then the joint likelihood in this set up is

$$L(x_{(1)}, x_{(2)}, \dots, x_{(n)} | g, f) = \frac{r!(n-r)!}{\varphi_r(G, F)} \prod_{i=1}^r g(x_{(i)}) \prod_{i=r+1}^n f(x_{(i)})$$

If g and f are exponential as above, then

$$L(x_{(1)}, x_{(2)}, \dots, x_{(n)} | \phi, \theta) = \frac{r!(n-r)!}{\varphi_r(\phi, \theta)} \frac{1}{\theta^r} e^{-\sum_{i=1}^r x_{(i)}/\phi} \frac{1}{\theta^{n-r}} e^{-\sum_{i=r+1}^n x_{(i)}/\theta}$$

The likelihood equations are

$$\frac{\partial \ln L}{\partial \phi} = \frac{\partial}{\partial \phi} \ln \varphi_r(\phi, \theta) - \frac{r}{\phi} + \frac{\sum_{i=1}^r x_i}{\phi^2}$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{\partial}{\partial \theta} \ln \varphi_r(\phi, \theta) - \frac{n-r}{\theta} + \frac{\sum_{i=r+1}^n x_i}{\theta^2}$$

An example (simulation):

- $n=15$, If $g \sim \text{exp}(0.04)$ and $f \sim \text{exp}(5)$, $\phi \ll \theta$,

The observations are

0.01339, 0.02679, 0.03442, 0.05519,
0.09459, 0.32254, 0.64367, 1.19427,
3.00276, 3.14612, 3.15643, 3.94635,
5.17659, 9.79405 and 12.52736

The detection of inliers is done as follows:

- evaluate for each fixed r , the maximum likelihood say \hat{L}_r , and then consider \hat{r} being that value of r for which the likelihood is maximum.

for the above data set the likelihood plot and the psi function ($\varphi_r(\phi, \theta)$) are given in Figures 1 and 2 respectively.

Fig.1.
Likelihood plot

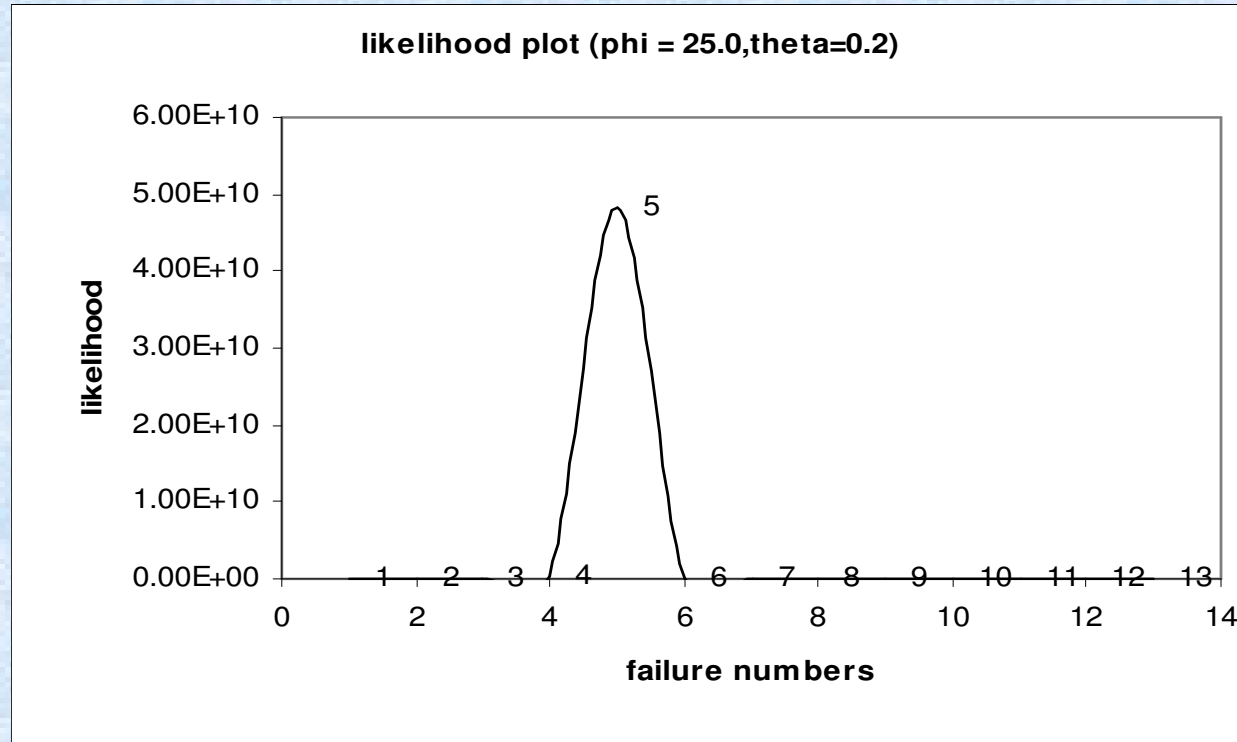
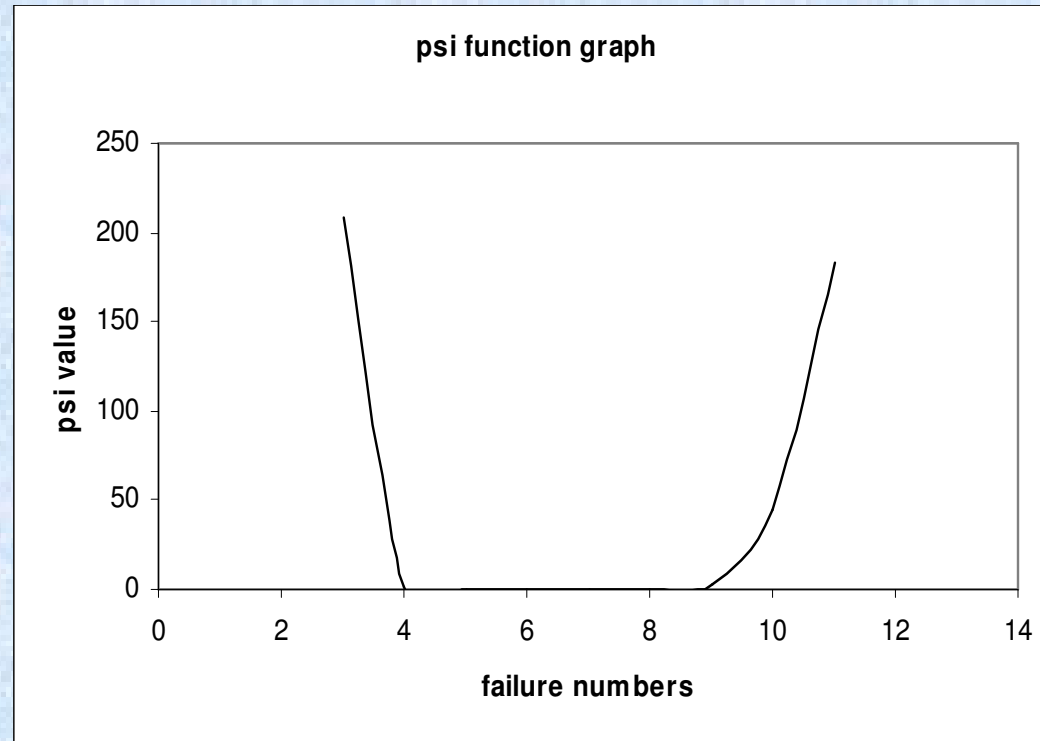


Fig.2.
Psi function graph



The estimates are :

$$\hat{\phi} = 18.64805 \text{ with mean life as } 0.05364$$

$$\hat{\theta} = 0.215466 \text{ with mean life as } 4.6411$$

$$\hat{r} = 5, \text{ where the likelihood is maximum}$$

$$\varphi_r(\hat{\phi}, \hat{\theta}) = 1.46\text{E-}13$$

Inliers as instantaneous and early failures

Assume that the data is usually consisting of r_0 instantaneous failures, r_1 early failures and the rest belongs to target population. Then

(i) **under identified inliers model the likelihood is**

$$L = \binom{n}{r_0} (1-p)^{r_0} p^{n-r_0} \left\{ \prod_{i=1}^{r_1} g(x_i, \phi) \prod_{i=r_1}^n f(x_i, \theta) \right\}$$

where p is the proportion of positive observations.

(ii). Under Labeled slippage inliers model, the likelihood is

$$L = \binom{n}{r_0} (1-p)^{r_0} p^{n-r_0} \left\{ \frac{r_1!(n-r_0-r_1)!}{\varphi_r(G, F)} \prod_{i=1}^{r_1} g(x_i, \phi) \prod_{i=r_1}^n f(x_i, \theta) \right\}$$

The above likelihood of the sample assumes that between the experiments when units are placed on test, we do not know which of the units fail instantaneously i.e. $X_{i_1} = 0, X_{i_2} = 0, \dots, X_{i_{r_0}} = 0$, and which fail early, i.e. those units whose failure time distribution is $g(x)$, with failure rate much larger than that of the failure time distribution of the target distribution whose failure rate is considerably smaller.

Suppose $g(x)$ and $f(x)$ are exponential in their canonical form, then

$$\begin{aligned}
 \varphi_r(\phi, \theta) &= \int_{-\infty}^{\infty} (n - r_0 - r_1)\theta [1 - e^{-\phi x}]^{r_1} e^{-(n-r_1)\theta x} dx \\
 &= \frac{(n - r_0 - r_1)\theta}{\phi} B\left(r_1 + 1, \frac{(n - r_0 - r_1)\theta}{\phi}\right), \\
 &= \frac{(n - r_0 - r_1)\theta}{\phi} \frac{r_1! \Gamma\left(\frac{(n - r_0 - r_1)\theta}{\phi}\right)}{\Gamma\left(\frac{(n - r_0 - r_1)\theta}{\phi} + r_1 + 1\right)}
 \end{aligned}$$

Therefore ,

$$\ln \varphi_r(\phi, \theta) = C + \ln \theta - \ln \phi + \ln \Gamma(z) - \ln \Gamma(z + r_1 + 1),$$

- Where $z = \frac{(n - r_0 - r_1)\theta}{\phi}$. Therefore,

$$\begin{aligned} \frac{\partial}{\partial \phi} \ln \varphi_{r_1}(\phi, \theta) &= -\frac{1}{\phi} + \frac{\partial}{\partial \phi} \ln \Gamma(z) \frac{dz}{d\phi} - \frac{\partial}{\partial \phi} \ln \Gamma(z + r_1 + 1) \frac{dz}{d\phi} \\ &= -\frac{1}{\phi} + [\Psi(z) - \Psi(z + r_1 + 1)] \left[-\frac{(n - r_0 - r_1)\theta}{\phi^2} \right], \end{aligned}$$

where $\Psi(z) = \frac{\partial}{\partial z} \ln \Gamma(z)$. And

$$\begin{aligned} \frac{\partial}{\partial \phi} \ln \varphi_{r_1}(\phi, \theta) &= \frac{1}{\theta} + \frac{\partial}{\partial \theta} \ln \Gamma(z) \frac{dz}{d\theta} - \frac{\partial}{\partial \theta} \ln \Gamma(z + r_1 + 1) \frac{dz}{d\theta} \\ &= \frac{1}{\theta} + [\Psi(z) - \Psi(z + r_1 + 1)] \left[\frac{(n - r_0 - r_1)}{\phi} \right] \end{aligned}$$

- Using Abramovitz and Stegun (1965), we get

$$\Psi(z) - \Psi(z + r_1 + 1) = -\sum_{j=1}^{r_1} \frac{1}{z + j}.$$

- Hence the likelihood equations are

$$\frac{\partial}{\partial \phi} \ln L = \frac{r_1 + 1}{\phi} - \frac{(n - r_0 - r_1)\theta}{\phi^2} \sum_{j=1}^{r_1} \frac{1}{z + j} - \sum_{i=1}^{r_1} x_{(i)} = 0$$

and

$$\frac{\partial}{\partial \theta} \ln L = \frac{(n - r_0 - r_1 - 1)}{\theta} - \frac{(n - r_0 - r_1)}{\phi} \sum_{j=1}^{r_1} \frac{1}{z + j} - \sum_{i=r_1+1}^{n-r_0} x_{(i)} = 0$$

Another example

This example is based on Vanmann's (1995) data on drying of woods under different schedules. The data on Experiment 2 on two batches of 37 boards by using two different schedules are:

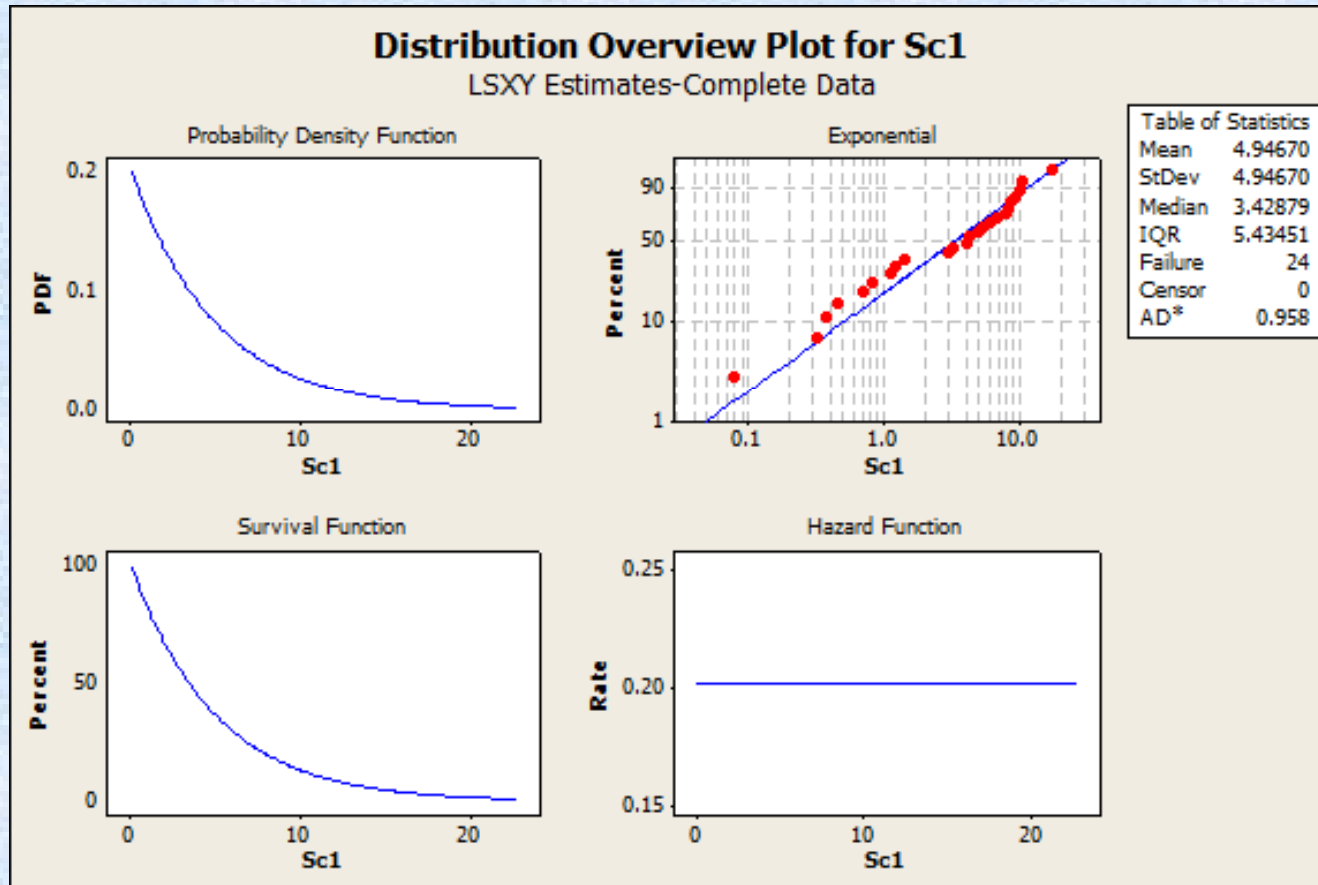
Schedule 1. $x_i = 0$, $i=1,2,\dots,13$ and the other positive observations arranged in increasing order of their magnitude are 0.08, 0.32, 0.38, 0.46, 0.71, 0.82, 1.15, 1.23, 1.40, 3.00, 3.23, 4.03, 4.20, 5.04, 5.36, 6.12, 6.79, 7.90, 8.27, 8.62, 9.50, 10.15, 10.58 and 17.49.

Schedule 2. $x_i = 0$, $i=1,2,\dots,17$ and the other positive observations arranged in increasing order of their magnitude are 0.02, 0.02, 0.02, 0.04, 0.09, 0.23, 0.26, 0.37, 0.93, 0.94, 1.02, 2.23, 2.79, 3.93, 4.47, 5.12, 5.19, 5.39, 6.83 and 8.22.

Goodness of Fit Test (positive observations)

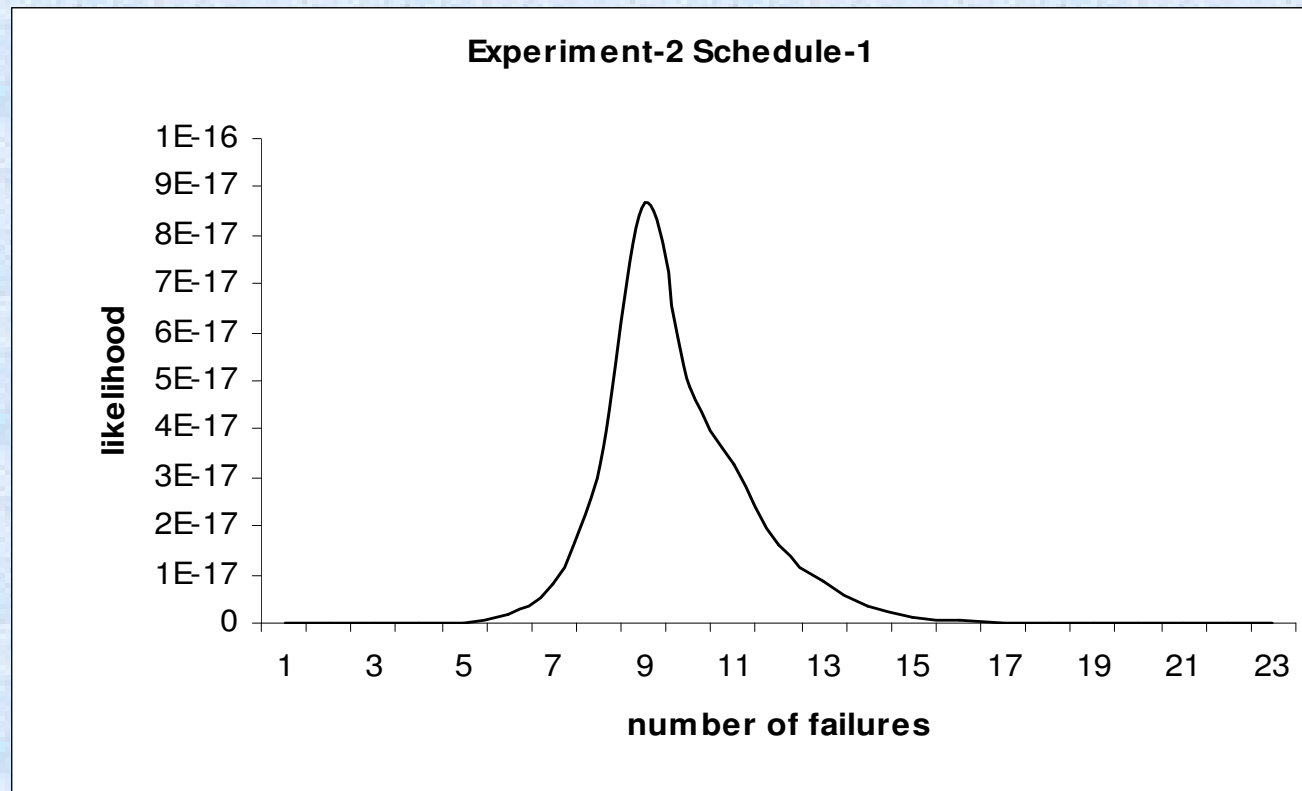
<i>Distribution</i>	<i>AD</i>	<i>P</i>	<i>LRT P</i>
<i>Normal</i>	0.572	0.109	
<i>Lognormal</i>	0.592	0.096	
3-Parameter Lognormal	0.544	*	0.728
Exponential	0.452	0.523	
2-Parameter Exponential	0.424	>0.250	0.286
Weibull	0.504	0.195	
3-Parameter Weibull	0.676	0.084	0.127
Smallest Extreme Value	0.708	0.054	
Largest Extreme Value	0.548	0.153	
Gamma	0.469	>0.250	
3-Parameter Gamma	0.629	*	0.057

Distribution overview plot of positive observations (Exponential)



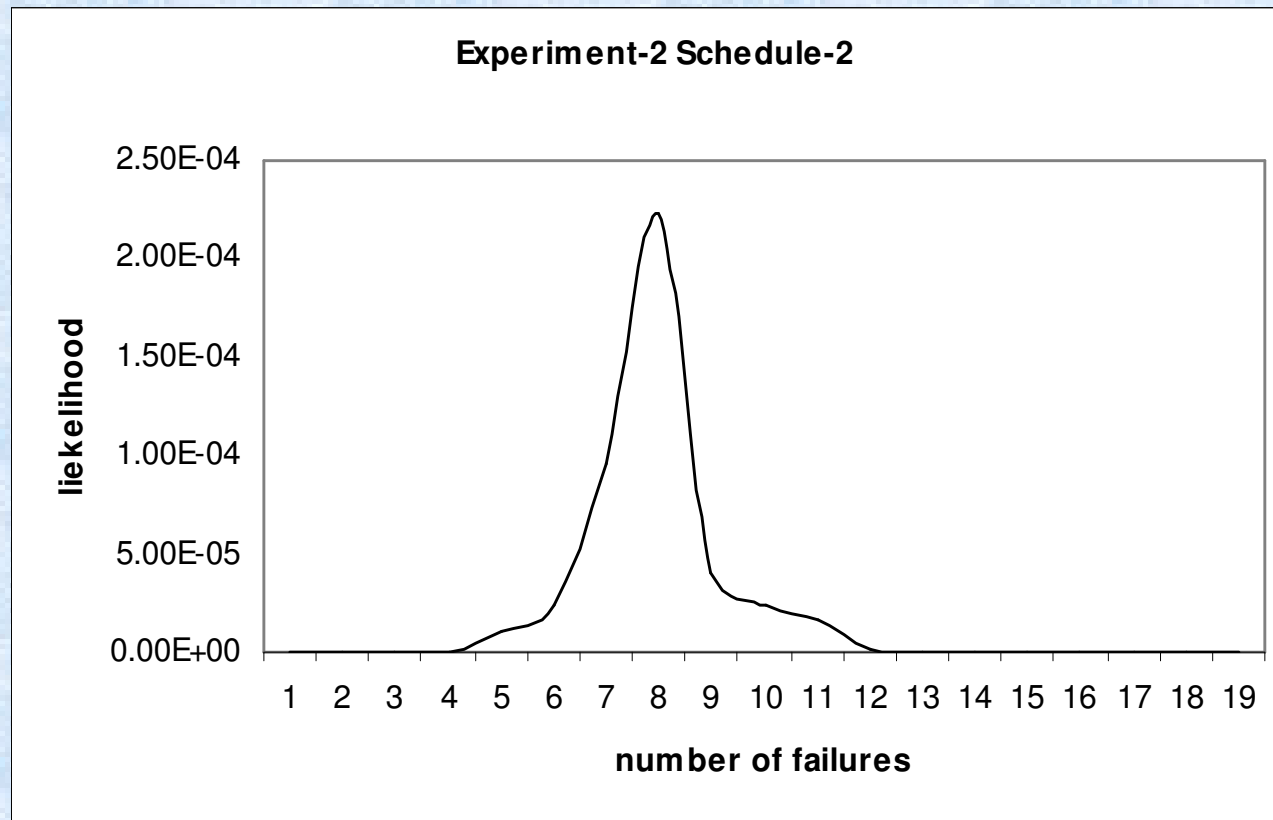
The likelihood plot under identified inliers model for Schedule-1 (all observations)

Fig. 3.



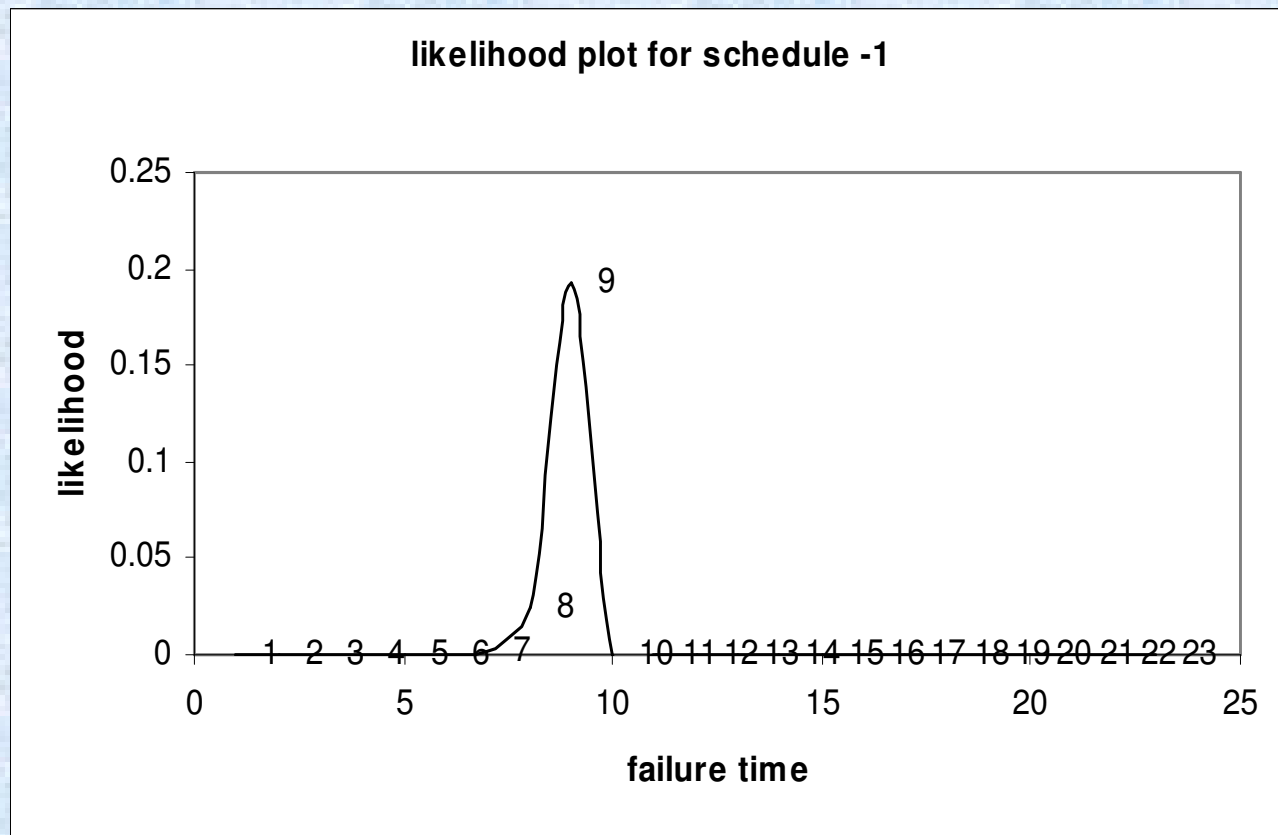
The likelihood plot under identified inliers model for Schedule-2

Fig.4.



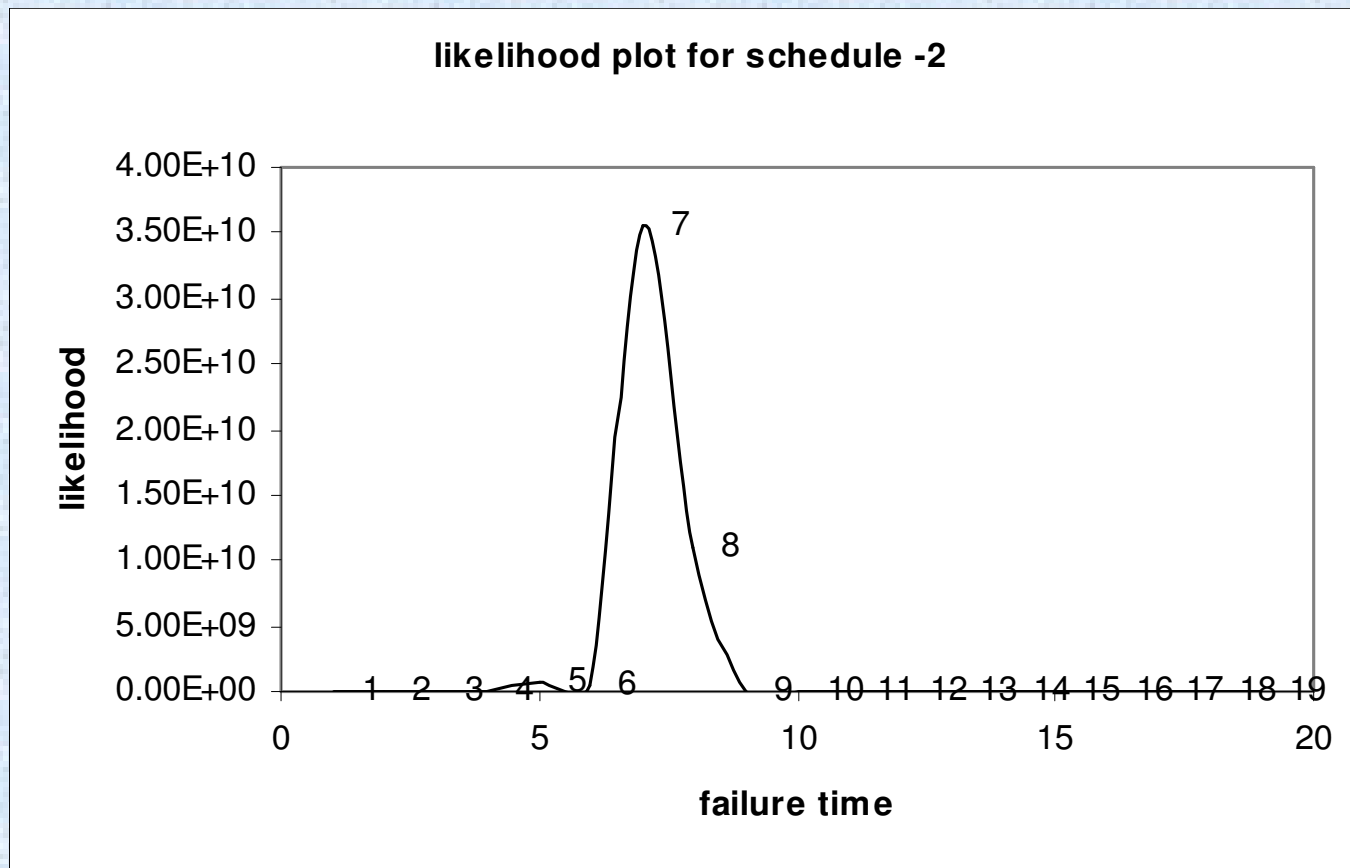
The likelihood plot under Labeled slippage inliers model for Schedule-1

Fig. 5.



The likelihood plot under Labeled slippage inliers model for Schedule-2

Fig.6.



The estimates under labeled slippage inliers model are:

For Schedule-1

- $\hat{r}_0 = 13$

- $\hat{r}_1 = 9$

- $\hat{\phi} = 0.97287$

- $\hat{\theta} = 6.37918$

for Schedule-2

- $\hat{r}_0 = 17$

- $\hat{r}_1 = 7$

- $\hat{\phi} = 0.12046$

- $\hat{\theta} = 3.67901$

Inliers detection using Schwarz information criterion (SIC)

(Muralidharan and Kale, 2008, JRSS)

Denoting the expectation of X_i by λ_i , $i=1,2,\dots,n$, we consider the following model of no inliers in the model as

$$x_i \quad \text{Model}(0) : \lambda_i = \theta, i=1,2,\dots,n$$

And the model with r inliers as

$$\text{Model}(r) : \lambda_i = \begin{cases} \phi, 1 \leq i \leq r \\ \theta, r+1 \leq i \leq n \end{cases}$$

The SIC scheme :Schwarz (1978)

- $SIC = -2\log L(\hat{\Theta}) + p \log(n),$
- where

$L(\hat{\Theta})$: the maximum of likelihood function
and p is the number of free parameters
that need to be estimated under the model.

Procedure:

model(0) is selected with no inliers if

$$SIC(0) \leq \min_{1 \leq r \leq n-1} SIC(r)$$

the model(r) is selected if

$$SIC(0) > \min_{1 \leq r \leq n-1} SIC(r)$$

Power of SIC procedure.

θ/ϕ r	4	6	8	10	12	14
3	0.509	0.785	0.896	0.946	0.982	1.00
4	0.514	0.822	0.934	0.968	0.996	1.00
5	0.555	0.874	0.957	0.999	1.00	1.00
6	0.586	0.887	0.988	1.00	1.00	1.00

Vanmann's (1995) data

Schedule 1. $x_i = 0$, $i=1,2,\dots,13$ and the other positive observations arranged in increasing order of their magnitude are

0.08, 0.32, 0.38, 0.46, 0.71, 0.82, 1.15,
1.23, 1.40, 3.00, 3.23, 4.03, 4.20, 5.04,
5.36, 6.12, 6.79, 7.90, 8.27, 8.62, 9.50,
10.15, 10.58 and 17.49.

The computed value of $SIC(0)= 127.1460$ and

the corresponding $SIC(r)$'s are

- 124.0335, 121.2333, 118.0726, 115.1593, 113.3191, 111.5499, 110.6135, 109.4866, 108.4856, 110.4540, 111.7338, 113.3368, 114.4581, 115.8601, 117.0348, 118.3385, 119.6651, 121.2188, 122.5813, 123.7872, 125.0636, 126.2976, 127.3799.
- Clearly, $SIC(0)=127.1460 > SIC(9)=\min_{1 \leq r \leq n-1} SIC(r) = 108.4856$.
Hence $\hat{r} = 9$.

Testing of Hypothesis

Tests for detecting inliers

1. Likelihood Ratio Test:

$H_0 : X_{(1)}, X_{(2)}, \dots, X_{(n)}$ are order statistics from F

$H_1 : X_{(1)}, X_{(2)}, \dots, X_{(r)}$ are order statistics from G and

$X_{(r+1)}, X_{(r+2)}, \dots, X_{(n)}$ are order statistics from F

Test Statistics:

$$\phi(x) = \begin{cases} 1, & \prod_{i=1}^r \frac{g_0(x_i)}{f_0(x_i)} > C_\alpha \\ 0, & \text{otherwise} \end{cases}$$

where α is such that $E[\phi(X) | H_0] = \alpha$

Theorem. Under the labeled slippage

alternative, $\frac{X_{(1)}}{X_{(i)}} \xrightarrow{p} 0$, as, $i=k+1, k+2, \dots, n$.

Proof: the joint density of $X_{(1)}$ and $X_{(k+1)}$ is

$$f(x_{(1)}, x_{(k+1)}) = \frac{(n-k)k\lambda}{\varphi(1,2,\dots,k)} e^{-\lambda x_{(1)}} \left[e^{-\lambda x_{(1)}} - e^{-\lambda x_{(k+1)}} \right]^{k-1} e^{-(n-k)x_{(k+1)}}$$

Hence for all $a \in (0, 1)$, we get

$$\begin{aligned} & P\left[\frac{X_{(1)}}{X_{(k+1)}} < a \mid H_{LK}\right] \\ &= \frac{(n-k)k\lambda}{\varphi(1,2,\dots,k)} \sum_{j=0}^{k-1} \frac{(-1)^j \binom{k-1}{j}}{[\lambda(k-1-j) + (n-k)] \left\{ \frac{1}{a} [\lambda(k-1-j) + (n-k)] + \lambda(j+1) \right\}} \end{aligned}$$

Thus we get $\frac{X_{(1)}}{X_{(k+1)}} \xrightarrow{p} 0$ as $\lambda \rightarrow \infty$. The proof is

Completed by noting that $0 \leq \frac{X_{(1)}}{X_{(i)}} \leq \frac{X_{(1)}}{X_{(k+1)}} \Rightarrow X_{(k+1)} < X_{(i)}$

, $i=k+2, \dots, n$, which proves the theorem.

2. Dixon's Test:

Test Statistic:

$$\phi(x) = \begin{cases} 1, & \frac{X_{(2)}}{X_{(1)}} > d \\ 0, & \text{otherwise} \end{cases}$$

where

$$d = \frac{n - \alpha}{(n - 1)\alpha}$$

The power of the test is

$$P_k(\lambda) = \frac{k\lambda + n - k}{\lambda[1 + (k - 1)d] + (n - k)d}, \lambda = \frac{\phi}{\theta}$$

The values of $P_1(\lambda)$ and $P_k(\lambda)$ for $\alpha = 0.05$

n	k	$\lambda=1$	$\lambda=5$	$\lambda=10$	$\lambda=15$	$\lambda=20$	$\lambda=30$
10	1	.050	.069	.091	.112	.132	.171
	2	.050	.062	.069	.073	.075	.078
	4	.050	.056	.057	.058	.058	.059
20	1	.050	.059	.071	.082	.093	.114
	2	.050	.057	.064	.068	.071	.075
	4	.050	.055	.057	.058	.059	.060
	6	.050	.053	.055	.056	.055	.056
30	1	.049	.056	.064	.072	.079	.074
	2	.049	.055	.060	.064	.067	.072
	4	.049	.054	.057	.058	.059	.060
	6	.049	.053	.055	.056	.056	.056

2. Cochran's Test:

Test Statistic:

$$\phi(x) = \begin{cases} 1, & \frac{\sum X_{(i)}}{X_{(1)}} > d \\ 0, & \text{otherwise} \end{cases}$$

where

$$d = \frac{n}{1 - (1 - \alpha)^{1/(n-1)}} - 1$$

The power of the test for one inlier is

$$P_1(\lambda) = 1 - \left(\frac{d - n + 1}{d + \lambda} \right)^{n-1}, \lambda = \frac{\phi}{\theta}$$

The values of $P_1(\lambda)$ and $P_k(\lambda)$ for $\alpha=0.05$

n	k	$\lambda=1$	$\lambda=5$	$\lambda=10$	$\lambda=15$	$\lambda=20$	$\lambda=30$
10	1	.050	.069	.093	.116	.138	.179
	2	.050	.052	.079	.093	.105	.148
	4	.050	.046	.067	.078	.088	.089
20	1	.050	.059	.072	.083	.095	.118
	2	.050	.057	.067	.078	.079	.095
	4	.050	.055	.057	.068	.068	.070
	6	.050	.052	.054	.054	.057	.066
30	1	.049	.056	.065	.073	.080	.098
	2	.049	.055	.060	.065	.077	.082
	4	.049	.054	.058	.060	.069	.074
	6	.049	.053	.056	.058	.060	.065

Masking effect

Let X_1, X_2, \dots, X_n be n observations

H_0 : all X 's are from F

H_1 : the (inlier) discordant observations are from G .

Let $T(X)$ be a test statistic to detect a single discordant observation, with critical region, say $C_{n,\alpha}$

- Due to lack of information about the number of discordant observations present in the sample, however, the true situation may not be specified by H_1 completely and more than one discordant observation may be present in the sample.
- In such cases, a test statistic $T(X)$ suggested for detection of a single discordant observation, may fail to detect a single inlier as discordant even when it is discordant, or the probability of detection of a discordant observation may decrease when additional observations are present in the sample.
- Such a phenomenon is called the masking effect.
- The masking effect has been widely discussed by various authors and some of the recent references are Barnett and Lewis (1978), Hawkins (1980), David (1981) and Bendre and Kale (1985) among others.

Thus the masking effect

$$M_{\lambda} = P_1(\lambda) - P_2(\lambda)$$

where

$$P_1(\lambda) = P(T(X) \in C_{n,\alpha} \mid H_1)$$

and

$$P_2(\lambda) = P(T(X) \in C_{n,\alpha} \mid \text{more than one observation follow df } G)$$

The limiting masking effect

$$M = \lim_{\lambda \rightarrow \lambda_0} M_{\lambda}$$

A test is said to suffer from masking effect if the measure M is positive.

and is said to be free from masking effect if M is zero.

For a consistent test, $\lim_{\lambda \rightarrow \lambda_0} M_{\lambda} = 1$ and for such a test $M \geq 0$.

CONCLUSION

- Instantaneous and early failures are together treated as inliers.
- Different models are considered.
- Among the models Labeled slippage and identified inliers model performs equally well.
- Inliers form a group of observations
- After detection inliers cannot be discarded from the experiment as done in outliers theory.

Present work

- Inlier(s) and outlier(s) detection simultaneously
- Different $g(x)$ and $f(x)$ in the model
- Sequential estimation and Bayesian estimation completed
- Other estimation procedures

Discussions and Suggestions

References

- Abramovitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical functions.*, General publishing Company Ltd, Canada.
- Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability Mass at the origin. *J. Amer. Stat. Assn.* Vol. 50, 901-908.
- Barnett, V. and Lewis, T. (1984). *Outliers in Statistical data*, John Wiley & Sons, New York.
- Gather, U. and Kale, B.K (1986). *Maximum likelihood estimation in the Presence of outliers.*, Pre-print 86-951. Dept. of statistics, Iowa State University, Ames, Iowa.

References.....

- Jayade, V.P. and Prasad, M.S. (1990). Estimation of parameters of mixed failure time distribution. *Comm. Statist. – Theory and Methods*, 19(12), 4667-4677.
- Kale, B.K. (2001). *Industrial Mathematics and Statistics – Chapter 20*, Narosa Publishers, New Delhi.
- Kale, B.K. and Muralidharan, K. (2000). Optimal estimating equations in mixture Distributions accommodating instantaneous or early failures. *Journal of Indian Statistical Assoc.*, 38, 317-329.
- Kleyle , R.M. and Dahiya, R.L. (1975). Estimation of parameters of mixed failure time distribution from censored data. *Comm. Statist. – Theory and Methods*, 4(9),873-882.

References.....

- Muralidharan, K. (1999). Tests for the mixing proportion in the mixture of a degene-rate and exponential distribution. *J. Indian Stat. Assn.*, Vol. 37, issue 2.
- Muralidharan, K. (2000). The UMVUE and Bayes estimate of reliability of mixed failure time distribution., *Comm. Statist- Simulations & Computations*, 29(2), 603-619.
- Muralidharan, K. (2005). Estimation in presence of early failures. *Statistical Methods* Vol. 7(2), 81-95.
- Muralidharan, K. and Kale, B.K. (2002). Modified Gamma distribution with Singularity at zero. *Comm. Statist- Simulations & computations*, 31(1), 143-158.
- Muralidharan, K. and Kale, B. K. (2005). Inliers detection using schwarz information critetion. *JRSS*.

References.....

- Muralidharan and Lathika (2003). Optimal estimating equations in zero inflated generalized poisson distribution, *Far East Journal of Theoretical Statistics*, Vol. 10, No. 2., 123-131.
- Muralidharan and Lathika (2004). A note on variance-covariance matrix of Weibull distribution, *Journal of Indian Statistical Association*, Vol. 42, No. 1, 75-78.
- Muralidharan and Lathika (2005). Statistical modeling of rainfall data using modified Weibull distribution., *Mausam, Indian Journal of Meteorology, Hydrology and Geophysics*. Vol. 56, No.4, 765-770.

References.....

Muralidharan and Lathika (2004). The concept of inliers.

Proceedings of first Sino- International Symposium on Probability, Statistics and Quantitative Management., Taiwan, October, 77-92.

Muralidharan and Lathika (2005). Some inferences on mixture of rayleigh and exponential distribution,

Proceedings of the second Sino-International Symposium on Probability, Statistics and Quantitative Management, Taiwan, October, 189-200.

References

- Kale, B.K. and Muralidharan, K. (2007). Masking effect of inliers. *Journal of Indian Statist. Assoc.* 45(1), 33-49.
- Kale, B.K. and Muralidharan, K. (2006). maximum likelihood estimation in presence of Inliers . *Journal of Indian Society for Prob. and Statist.* 10, 65-80.
- Lai, C. D., Khoo, B. C., Muralidharan, K. and Xie, M. (2007). Weibull model allowing nearly instantaneous failures. *J. Applied Mathematics and Decision Sciences*, Article ID 90842, 11 pages.

References

- Vannman. K. (1991). *Comparing samples from nonstandard mixtures of distributions with Applications to quality comparison of wood.* Research report 1991:2 submitted to Division of Quality Technology, Lulea University, Lulea, Swedon.
- Vannman, K. (1995). On the distribution of the estimated mean from the nonstandard Mixtures Of distribution. *Comm. Statist. –Theory and Methods*, 24(6), 1569-1584.



THANK YOU