

AD-A057 871

SYRACUSE UNIV N Y

BAYESIAN SOFTWARE PREDICTION MODELS. VOLUME II. CLASSICAL AND B--ETC(U)

JUL 78 K OKUMOTO, A L GOEL

F30602-76-C-0097

F/G 9/2

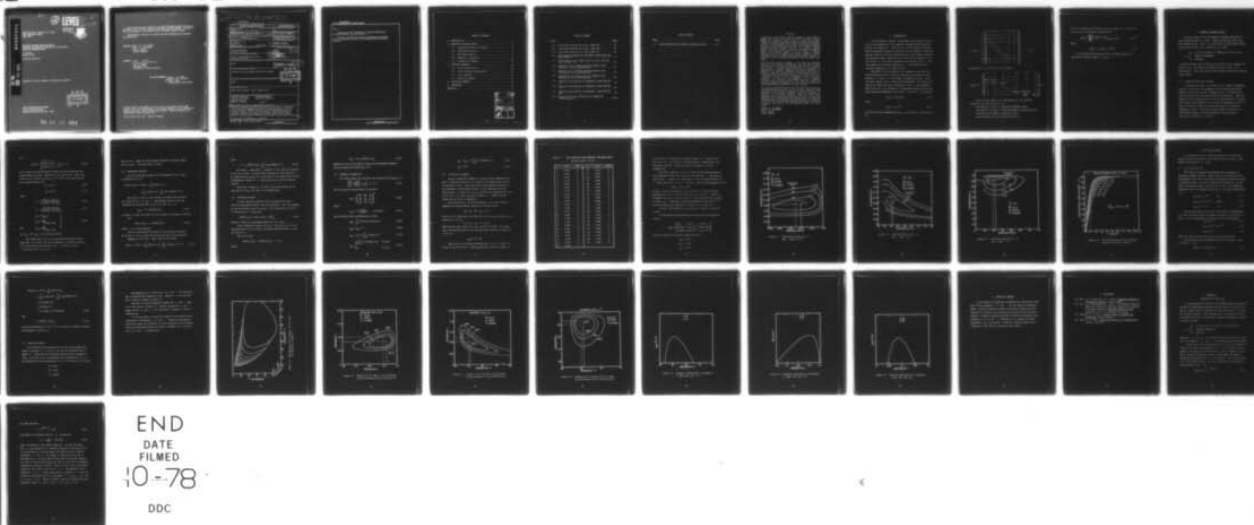
UNCLASSIFIED

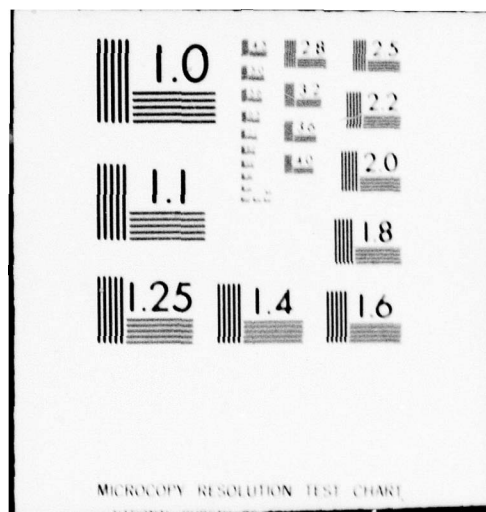
TR-78-2

RADC-TR-78-155-VOL-2

NL

1 OF 1
AD
A057 871





ADA057871

AD NO.

DDC FILE COPY

②

LEVEL

III 2
A057870

A057872
A057873



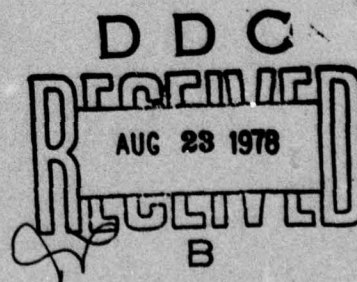
RADC-TR-78-155, Volume II (of five)
Final Technical Report
July 1978

BAYESIAN SOFTWARE PREDICTION MODELS
Classical and Bayesian Inference for the Software
Imperfect Debugging Model

K. Okumoto
Amrit L. Goel

Syracuse University

Approved for public release; distribution unlimited.



ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441

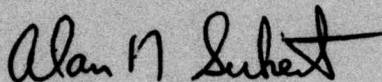
78 08 22 061

111

This report has been reviewed by the RADC Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

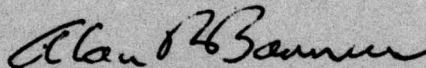
RADC-TR-78-155, Volume II (of five) has been reviewed and is approved for publication.

APPROVED:



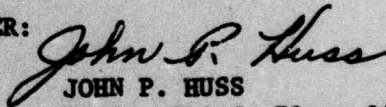
ALAN N. SUKERT
Project Engineer

APPROVED:



ALAN R. BARNUM
Assistant Chief
Information Sciences Division

FOR THE COMMANDER:



JOHN P. HUSS
Acting Chief, Plans Office

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (ISIS) Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return this copy. Retain or destroy.

(19)

TR-78-155-VOL-2

(14)

TR-78-2

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-78-155, Vol II (of five)	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER (9)
4. TITLE (and Subtitle) BAYESIAN SOFTWARE PREDICTION MODELS, Volume II. Classical and Bayesian Inference for the Software Imperfect Debugging Model.		5. REPORT DATE Final Technical Report Dec 75 - Mar 78
7. AUTHOR(s) K. Okumoto Amrit L. Goel		6. PERFORMING ORG. REPORT NUMBER Technical Report No. 78-2
9. PERFORMING ORGANIZATION NAME AND ADDRESS Syracuse University Syracuse NY 13210		8. CONTRACT OR GRANT NUMBER(s) F30602-76-C-0097
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (ISIS) Griffiss AFB NY 13441		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBER 62702F 55811403
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Same		12. REPORT DATE July 78
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		13. NUMBER OF PAGES 38
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) Same		15. SECURITY CLASS. (of this report) UNCLASSIFIED
18. SUPPLEMENTARY NOTES RADC Project Engineer: Alan N. Sukert (ISIS)		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Imperfect Debugging Confidence Regions Statistical Inference Asymptotic Properties Maximum Likelihood Bayesian Inference <i>lambda</i>		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This report presents two methods for statistical inference of the parameters of the imperfect debugging model proposed by Goel and Okumoto. Using the method of maximum likelihood, the mle's, the likelihood contours and the confidence regions for N, p and λ are obtained. A Bayesian approach is presented to obtain the Bayesian point estimates and the H.P.D. regions. Numerical examples based on simulated data are used for illustrative purposes. Volume V will be published at a later date.		

DDC
RECEIVED
AUG 23 1978
B

DD FORM 1473

JAN 73

EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

339 600

Liu

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

Block 7.

¹Research Assistant, Department of Industrial Engineering & Operations Research, Syracuse University.

²Professor, Department of Industrial Engineering & Operations Research, and School of Computer and Information Science, Syracuse University.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

TABLE OF CONTENTS

1. INTRODUCTION	1
2. MAXIMUM LIKELIHOOD METHOD.	4
2.1 Likelihood Function and MLE's.	4
2.2 Likelihood Contours.	8
2.3 Confidence Regions	9
2.4 Asymptotic Properties.	10
2.5 Illustrative Example	11
3. BAYESIAN INFERENCE	18
3.1 Prior Distributions.	18
3.2 Joint Posterior Distribution	19
3.3 H.P.D. Regions	20
3.4 Numerical Example.	21
4. CONCLUDING REMARKS	30
5. REFERENCES	31
APPENDIX A	32

SECTION 10	
NTD	NTD Section ☒
SEC	SEC Section ☐
QUALIFICATION	☐
AUTHORIZATION	
BY	
DISTRIBUTION/AVAILABILITY CODES	
Dist.	AVAIL and/or SPECIAL
A	

1 78 08 22 061

LIST OF FIGURES

Figure		Page
2.1	Likelihood Contours for p and λ when $N=\hat{N}$	14
2.2	Likelihood Contours for N and λ when $p=\hat{p}$	15
2.3	Likelihood Contours for N and p when $\lambda=\hat{\lambda}$	16
2.4	Relationship Between the Confidence Coefficient and the Constant ρ	17
3.1	90% Bayesian H.P.D. Region for N , p and λ for the data in Table 2.1.	23
3.2	Bayesian H.P.D. Regions with $N=\hat{N}$ Based on Non-informative Prior Distributions.	24
3.3	Bayesian H.P.D. Regions with $p=\hat{p}$ Based on Non-informative Prior Distributions.	25
3.4	Bayesian H.P.D. Regions with $\lambda=\hat{\lambda}$ Based on Non-informative Prior Distributions.	26
3.5	Posterior Distribution of Parameter N when $p=\hat{p}$ and $\lambda=\hat{\lambda}$	27
3.6	Posterior Distribution of Parameter p when $N=\hat{N}$ and $\lambda=\hat{\lambda}$	28
3.7	Posterior Distribution of Parameter λ when $N=\hat{N}$ and $p=\hat{p}$	29
A.1	Flow Chart of Data Simulation for Imperfect Debugging Model.	33-34

LIST OF TABLES

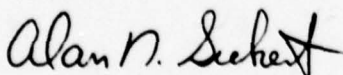
Table	Page
2.1 Data Simulated From Imperfect Debugging Model	12

EVALUATION

The necessity for more complex software systems in such areas as command and control and avionics has led to the desire for better methods for predicting software errors to insure that software produced is of higher quality and of lower cost. This desire has been expressed in numerous industry and Government sponsored conferences, as well as in documents such as the Joint Commanders' Software Reliability Working Group Report (Nov 1975). As a result, numerous efforts have been initiated to develop and validate mathematical models for predicting such quantities as the number of remaining errors in a software package, the time to achieve a desired reliability level, and a measure of the software reliability. However, early efforts have not produced models with the desired accuracy of prediction and with the necessary confidence limits for general model usage.

This effort was initiated in response to this need for developing better and more accurate software error prediction models and fits into the goals of RADC TPO No. 5, Software Cost Reduction (formerly RADC TPO No. 11, Software Sciences Technology), in the subthrust of Software Quality (Software Modeling). This report summarizes the development of classical and Bayesian estimates for parameters of a model for predicting quantities such as the expected number of remaining errors, achieved reliability, and time to detect and correct a specified number of errors that assumes a software error is not corrected at a given time with probability 1 (i.e. imperfect debugging). The importance of this development is that it represents the first attempt to develop software error prediction models that incorporate imperfect debugging, and thus more closely reflect the actual software error detection and correction process.

The theory and equations developed under this effort will lead to much needed predictive measures for use by software managers in more accurately tracking software development projects in terms of test time needed to achieve given reliability and error objectives. In addition, the associated confidence limits and other related statistical quantities developed under this effort will insure more widespread use of these modeling techniques. Finally, the predictive measures and equations developed under this effort will be applicable to current Air Force software development projects and thus help to produce the high quality, low cost software needed for today's systems.



ALAN N. SUKERT
Project Engineer

1. INTRODUCTION

In this report we present two methods for statistical inference of the parameters of the imperfect debugging model proposed by Goel and Okumoto [2]. The first one is the classical approach based on maximum likelihood estimation and the second is a Bayesian approach based on the prior distributions of the unknown parameters. The parameters under consideration are the initial number of software errors, N , the error occurrence rate for each error λ , and the probability of perfect debugging p . The probability of imperfect debugging is q where $q = 1 - p$.

The model in [2] is based on the assumption that the time between software errors follows an exponential distribution with parameter $i\lambda$ where i is the number of remaining errors. Also, the error removal time is taken to be negligible. By letting $X(t)$ denote the number of errors remaining at time t , the stochastic behavior of $X(t)$ is analyzed as a semi-Markov process and the one step transition probability from state i to state j is given by

$$Q_{ij}(t) = P_{ij} \cdot F_i(t), \quad (1.1)$$

where

$$F_i(t) = 1 - e^{-i\lambda t} \quad (1.2)$$

and the transition probabilities p_{ij} , $i, j = 0, 1, 2, \dots, N$ are given by

$$P = (P_{ij}) = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & \dots & i-1 & i & \dots & N-2 & N-1 & N \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ i-1 \\ i \\ \vdots \\ N-1 \\ N \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ p & q & 0 & \dots & \dots & \dots & 0 & 0 & 0 \\ 0 & p & q & \dots & \dots & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & p & q & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & \dots & \dots & p & q & 0 \\ 0 & 0 & \dots & \dots & \dots & \dots & 0 & p & q \end{bmatrix} \end{matrix} \quad (1.3)$$

Substituting (1.2) and (1.3) in (1.1) yields

$$\{Q_{ij}(t)\} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & \dots & N-2 & N-1 & N \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \\ N \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ pF_1(t) & qF_1(t) & 0 & \dots & 0 & 0 & 0 \\ 0 & pF_1(t) & qF_2(t) & \dots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & pF_{N-1}(t) & qF_{N-1}(t) & 0 \\ 0 & 0 & \dots & \dots & 0 & pF_N(t) & qF_N(t) \end{bmatrix} \end{matrix} \quad (1.4)$$

From the basic model (1.4), expressions for the following quantities have been derived in [2].

- Distribution of time to a completely debugged system.
- Distribution of time to a specified number of remaining errors.
- Distribution of number of remaining errors.
- Expected number of errors detected by time t .

Also, the reliability function at the k th stage, i.e. between the $(k-1)$ st and k th failures, is obtained as

$$R_k(x) = \sum_{j=0}^{k-1} \binom{k-1}{j} p^{k-j-1} q^j \bar{F}_{N-(k-j-1)}(x) \quad (1.5)$$

where

$$\bar{F}_N(x) = 1 - F_N(x) = e^{-N\lambda x}. \quad (1.6)$$

In the following sections we will use these results for statistical inference about N , p and λ .

2. MAXIMUM LIKELIHOOD METHOD

In this section we use the method of maximum likelihood to draw inferences about N , p and λ based on available data $(\underline{t}, \underline{y})$ for software errors. Here, $\underline{t}$ is the vector of times between software failures while $\underline{y}$ is a vector of y_i 's where

$$y_i = \begin{cases} 1, & \text{if the } i\text{th failure is caused by an error due to} \\ & \text{imperfect debugging,} \\ 0, & \text{otherwise.} \end{cases}$$

It should be noted that we make use of the data $(\underline{t}, \underline{y})$ because the process $X(t)$, the number of remaining errors at time t , is unobservable. Also, such data can be available from actual software error reports.

2.1 Likelihood Function and MLE's

As pointed out above, the state of $X(t)$ cannot be observed. However, we note that the sequence of error corrections forms a sequence of Bernoulli trials. Suppose that $(i-1)$ failures have been observed and the i th failure has not occurred yet. Then the number of errors eliminated up to now is distributed as a binomial distribution with parameters $(i-1, p)$ and its expectation is $p(i-1)$. Also, the expected number of errors occurred due to imperfect debugging is $q(i-1)$. Since the initial number of errors is N , the expected number of errors remaining in the software at this stage is given by $N - p(i-1)$.

Therefore, the distribution of the time between (i-1)st and ith failures is given by

$$f(t_i | N, p, \lambda) = [N - p(i-1)] \lambda e^{-[N - p(i-1)] \lambda t_i}, \quad i = 1, 2, \dots, n \quad (2.1)$$

where n is the number of observed software failures.

Then the likelihood function for a given $\underline{t}$ is

$$L_1(N, p, \lambda | \underline{t}) = \prod_{i=1}^n f(t_i | N, p, \lambda). \quad (2.2)$$

The next (ith) error will occur randomly from the remaining $[N - p(i-1)]$ errors and hence the probability of this error being due to the imperfect debugging category is $q(i-1)/[N - p(i-1)]$. Therefore, the distribution of Y_i is

$$P(Y_i = y_i | N, p) = \left\{ \frac{q(i-1)}{N - p(i-1)} \right\}^{y_i} \left\{ \frac{N - (i-1)}{N - p(i-1)} \right\}^{1-y_i}, \quad i = 1, 2, \dots, n \quad (2.3)$$

where

$$y_i = 0 \text{ or } 1.$$

Then, the likelihood function for given $\underline{y}$ is

$$L_2(N, p | \underline{y}) = \prod_{i=1}^n P(Y_i = y_i | N, p). \quad (2.4)$$

Due to the independence of $\underline{t}$ and $\underline{y}$, the likelihood function of N, p, λ can be written as

$$L(N, p, \lambda | \underline{t}, \underline{y}) = L_1(N, p, \lambda | \underline{t}) \cdot L_2(N, p, \lambda | \underline{y}). \quad (2.5)$$

Now we choose $\hat{N}$, $\hat{p}$ and $\hat{\lambda}$ which maximize (2.5). Maximizing $L(N, p, \lambda | \underline{t}, \underline{y})$ implies maximizing the log likelihood function. Let

$$l(N, p, \lambda | \underline{t}, \underline{y}) = \log L(N, p, \lambda | \underline{t}, \underline{y})$$

$$= n \log \lambda - \lambda \sum_{i=1}^n [N-p(i-1)] t_i + \sum_{i=1}^n y_i \log q(i-1) + \sum_{i=1}^n (1-y_i) \log [N-(i-1)] . \quad (2.6)$$

Then $\hat{N}$, $\hat{p}$ and $\hat{\lambda}$ must satisfy

$$\left. \begin{aligned} \frac{\partial l}{\partial N} &= 0 \\ \frac{\partial l}{\partial p} &= 0 \\ \frac{\partial l}{\partial \lambda} &= 0 \end{aligned} \right\} \quad (2.7)$$

or

$$\lambda \sum_{i=1}^n t_i = \sum_{i=1}^n \frac{1-y_i}{N-(i-1)} \quad (2.8)$$

$$\lambda \sum_{i=1}^n (i-1) t_i = \sum_{i=1}^n y_i / q \quad (2.9)$$

$$n/\lambda = \sum_{i=1}^n [N-p(i-1)] t_i . \quad (2.10)$$

Simultaneous non-linear equations (2.8), (2.9) and (2.10) can be solved by numerical methods as described below. From (2.10) we get

$$\lambda = n / \sum_{i=1}^n [N-p(i-1)] t_i . \quad (2.11)$$

Substituting (2.11) into (2.8) and (2.9), we get

$$f(N, p) = \sum_{i=1}^n \frac{1-y_i}{N-(i-1)} - \frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n [N-p(i-1)] t_i} = 0 , \quad (2.12)$$

and

$$\varphi(N, p) = \frac{qn \sum_{i=1}^n (i-1)t_i}{\sum_{i=1}^n \{N-p(i-1)\}t_i} - \sum_{i=1}^n y_i = 0. \quad (2.13)$$

We now apply the Newton-Raphson method [4] for solving the two simultaneous non-linear equations (2.12) and (2.13). Thus, for given initial values N_0 and p_0 of N and p the values of the first approximations are

$$N_1 = N_0 + h \quad (2.14)$$

$$p_1 = p_0 + k \quad (2.15)$$

where

$$h = - \frac{f_0 \varphi_{p,0} - \varphi_0 f_{p,0}}{f_{N,0} \varphi_{p,0} - \varphi_{N,0} f_{p,0}} \quad (2.16)$$

$$k = - \frac{f_{N,0} \varphi_0 - \varphi_{N,0} f_0}{f_{N,0} \varphi_{p,0} - \varphi_{N,0} f_{p,0}}, \quad (2.17)$$

$$\left. \begin{aligned} f_0 &\equiv f(N_0, p_0), \\ f_{N,0} &\equiv \left. \frac{\partial f}{\partial N} \right|_{N=N_0, p=p_0}, \\ f_{p,0} &\equiv \left. \frac{\partial f}{\partial p} \right|_{N=N_0, p=p_0} \end{aligned} \right\} \quad (2.18)$$

and

φ_0 , $\varphi_{N,0}$ and $\varphi_{p,0}$ are similarly defined.

The values of N and p are successively modified until equations (2.12) and (2.13) are satisfied to a defined accuracy; such values being the estimates $\hat{N}$, $\hat{p}$. Finally, we get $\hat{\lambda}$ by substituting $\hat{N}$ and $\hat{p}$

into (2.11). These are the maximum likelihood estimates (MLE's) of N , p and λ for given data $\underline{t}$ and $\underline{y}$.

2.2 Likelihood Contours

The log likelihood surface for the parameters N , p and λ is given by (2.6) as

$$\begin{aligned} l(N, p, \lambda | \underline{t}, \underline{y}) = & n \log \lambda - \lambda \sum_{i=1}^n (N - p(i-1)) t_i \\ & + \sum_{i=1}^n y_i \log q(i-1) + \sum_{i=1}^n (1 - y_i) \log \{N - (i-1)\}. \end{aligned}$$

For given $\underline{t}$ and $\underline{y}$, this defines a 4-dimensional surface as a function of N , p and λ . The maximum value of this log likelihood is obtained when $N = \hat{N}$, $p = \hat{p}$ and $\lambda = \hat{\lambda}$ i.e.

$$l_{\max} = \hat{l} = l(\hat{N}, \hat{p}, \hat{\lambda} | \underline{t}, \underline{y}). \quad (2.19)$$

In order to study the nature of this surface, we proceed as follows.

Let

$$l(N, p, \lambda | \underline{t}, \underline{y}) = \rho \cdot l(\hat{N}, \hat{p}, \hat{\lambda} | \underline{t}, \underline{y}), \quad (2.20)$$

where $\rho \geq 1$ is some constant.

We will investigate the nature of this surface by fixing $\hat{N}$, $\hat{p}$ and $\hat{\lambda}$, one at-a-time and varying the other two parameters.

Suppose we fix $N = \hat{N}$. Then, from (2.6) we have

$$f(p, \lambda) = n \log \lambda - \lambda \sum_{i=1}^n [\hat{N} - p(i-1)] t_i + \sum_{i=1}^n y_i \log q(i-1) = C, \quad (2.21)$$

where

$$C = \rho \ell(\hat{N}, \hat{p}, \hat{\lambda} | \underline{t}, \underline{y}) - \sum_{i=1}^n (1-y_i) \log[\hat{N} - (i-1)] . \quad (2.22)$$

By fixing ρ and hence C , equation (2.21) gives one contour in the $(p-\lambda)$ plane. To draw these contours for each value of ρ , we choose several values of p and solve (2.21) numerically for the corresponding values of λ . These pairs (p, λ) give the desired contour.

Similarly, contours in the $(N-\lambda)$ and $(N-p)$ planes can be obtained by fixing $p = \hat{p}$ and $\lambda = \hat{\lambda}$, respectively.

2.3 Confidence Regions

In many instances interest lies in studying the joint $100(1-\alpha)\%$ confidence region for the parameters. For this purpose we use the property that for large n the likelihood ratio has a χ^2 -distribution. In our case,

$$\ell(\hat{N}, \hat{p}, \hat{\lambda} | \underline{t}, \underline{y}) - \ell(N, p, \lambda | \underline{t}, \underline{y}) = \frac{1}{2} \chi_{3, \alpha}^2 \quad (2.23)$$

defines a $100(1-\alpha)\%$ confidence region for N , p and λ .

Joint confidence regions for (p, λ) , (N, p) and (N, λ) can be obtained from (2.23) by using a numerical method similar to that of Section 2.2.

Now, by writing

$$\ell(N, p, \lambda | \underline{t}, \underline{y}) = \rho \ell(\hat{N}, \hat{p}, \hat{\lambda} | \underline{t}, \underline{y}), \quad \rho \geq 1,$$

we get

$$\chi^2_{3;\alpha} = 2(1-\alpha)l(\hat{N}, \hat{p}, \hat{\lambda} | \underline{t}, \underline{y}) . \quad (2.24)$$

Equation (2.24) can be used to study the relationship between α and the confidence coefficient $(1-\alpha)$.

2.4 Asymptotic Properties

For large sample size the mle's are normally distributed i.e.

$$\begin{pmatrix} \hat{N} \\ \hat{p} \\ \hat{\lambda} \end{pmatrix} \sim N \left(\begin{pmatrix} N \\ p \\ \lambda \end{pmatrix}, \Sigma_{\text{cov}} \right) \quad \text{as } n \rightarrow \infty . \quad (2.25)$$

The variance-covariance matrix is given by

$$\Sigma_{\text{cov}} = \begin{pmatrix} r_{NN} & r_{Np} & r_{N\lambda} \\ r_{pN} & r_{pp} & r_{p\lambda} \\ r_{\lambda N} & r_{\lambda p} & r_{\lambda\lambda} \end{pmatrix}^{-1} \quad (2.26)$$

where

$$r_{a,b} = -E \left(\frac{\partial^2 l}{\partial a \partial b} \right), \quad a, b = N, p, \lambda . \quad (2.27)$$

For the model under consideration, we have

$$r_{NN} = \sum_{i=1}^n 1/[N-(i-1)][N-p(i-1)] \quad (2.28)$$

$$r_{Np} = r_{pN} = 0 \quad (2.29)$$

$$r_{N\lambda} = r_{\lambda N} = \frac{1}{\lambda} \sum_{i=1}^n 1/[N-p(i-1)] \quad (2.30)$$

$$r_{pp} = \begin{cases} \frac{1}{q} \sum_{i=1}^n (i-1)/[N-p(i-1)] & \text{if } q \neq 0 \\ - & \text{if } q = 0 . \end{cases} \quad (2.31)$$

$$r_{p\lambda} = r_{\lambda p} = -\frac{1}{\lambda} \sum_{i=1}^n (i-1)/(N-p(i-1)) \quad (2.32)$$

$$r_{\lambda\lambda} = n/\lambda^2. \quad (2.33)$$

2.5 Illustrative Example

We use a numerical example to illustrate the computations of mle's, likelihood contours, etc. based on the expressions derived in Sections 2.1-2.4. Since data from actual software projects is not available in the desired format, we use simulated data for this purpose. A total of 45 (t_i, y_i) values were simulated for $N=50$, $p=0.9$ and $\lambda=0.1$ and are given in Table 2.1. Details of the simulation are given in Appendix A.

For this data set, we solve equations (2.12) and (2.13) by applying the Newton-Raphson method with initial values

$$N_0 = 46 \quad \text{and} \quad p_0 = 1.0.$$

After six (6) iterations, the solution of (2.12) and (2.13), with an accuracy of 10^{-3} , is

$$\hat{N} = 51.3 \quad \text{and} \quad \hat{p} = 0.919.$$

Substituting these values in (2.11), we get $\hat{\lambda} = 0.085$. For these mle's the maximum value of the log-likelihood function is given by (2.19) as

$$l_{\max} = \hat{l} = -16.$$

Now we get the likelihood contours for $p = 1.1, 1.3$ and 1.5 . First we fix $N = \hat{N} = 51.3$. Solving equation (2.21) for various p ,

TABLE 2.1 DATA SIMULATED FROM IMPERFECT DEBUGGING MODEL
($N = 50$, $p = 0.9$, $\lambda = 0.1$)

I	t(I)	y(I)	I	t(I)	y(I)
1	0.296	0	24	0.127	0
2	0.156	0	25	0.034	0
3	0.239	0	26	0.013	0
4	0.174	0	27	0.444	0
5	0.204	0	28	1.690	0
6	0.182	0	29	0.037	0
7	0.323	0	30	0.034	0
8	0.174	0	31	0.142	0
9	0.365	0	32	0.287	0
10	0.074	0	33	0.568	0
11	0.087	0	34	1.310	0
12	0.230	0	35	1.668	1
13	0.520	0	36	0.754	1
14	0.084	0	37	1.451	0
15	0.380	0	38	0.038	1
16	0.114	0	39	0.499	1
17	0.396	0	40	0.372	0
18	0.256	0	41	0.058	0
19	0.200	0	42	0.529	0
20	0.072	0	43	0.359	0
21	1.253	0	44	1.020	0
22	0.518	0	45	2.083	0
23	0.904	0			

the contours are obtained as shown in Figure 2.1. Contours with $p = \hat{p}$ and $\lambda = \hat{\lambda}$ in the $(N-\lambda)$ and $(N-p)$ planes, respectively, are obtained similarly. These are shown in Figures 2.2 and 2.3, respectively.

Now we use equation (2.24) to study the relationship between the confidence coefficient $(1-\alpha)$ and the constant ρ . For given $\hat{t}$, and various ρ values, the coefficients $(1-\alpha)$ are obtained from the χ^2 table such that (2.24) is satisfied. Thus, for our example we have

$$\chi^2_{3;\alpha} = 2(1-\rho)(-16)$$

and for $\rho = 1.1$, the value of $(1-\alpha)$ from the χ^2 table is 0.638. Similarly for $\rho = 1.3$, $1-\alpha = .022$ and for $\rho = 1.5$, $1-\alpha = .01$. Plots of confidence level vs ρ for $\hat{t} = -10, -12, -14, -16, -18$ and -20 are given in Figure 2.4. Confidence levels corresponding to the value of ρ are also shown on the contours in Figures 2.1, 2.2 and 2.3.

Finally, the asymptotic distribution of $(\hat{N}, \hat{p}, \hat{\lambda})$ is given by (2.25).

The estimated variance-covariance matrix for the simulated data is

$$\hat{\Sigma}_{\text{cov}} = \begin{bmatrix} 35.5 & -0.0122 & -0.0128 \\ -0.0122 & 2.25 \times 10^{-3} & 6.08 \times 10^{-4} \\ -0.0128 & 6.08 \times 10^{-4} & 6.41 \times 10^{-4} \end{bmatrix}$$

and the estimated correlation coefficients are

$$\hat{\rho}_{Np} = -0.15$$

$$\hat{\rho}_{N\lambda} = -0.56$$

$$\hat{\rho}_{p\lambda} = -0.51$$

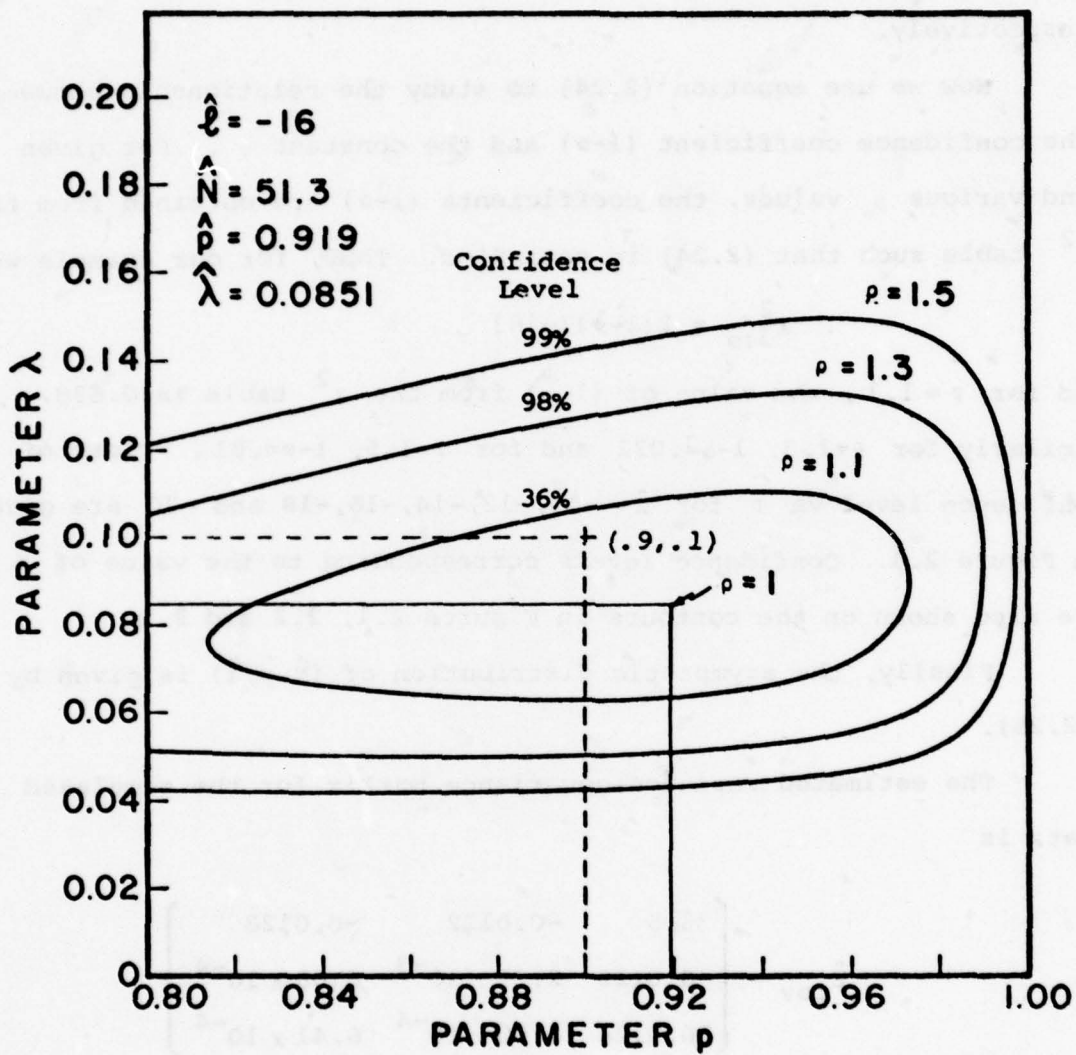


Figure 2.1 Likelihood Contours for p and λ when $N = \hat{N}$

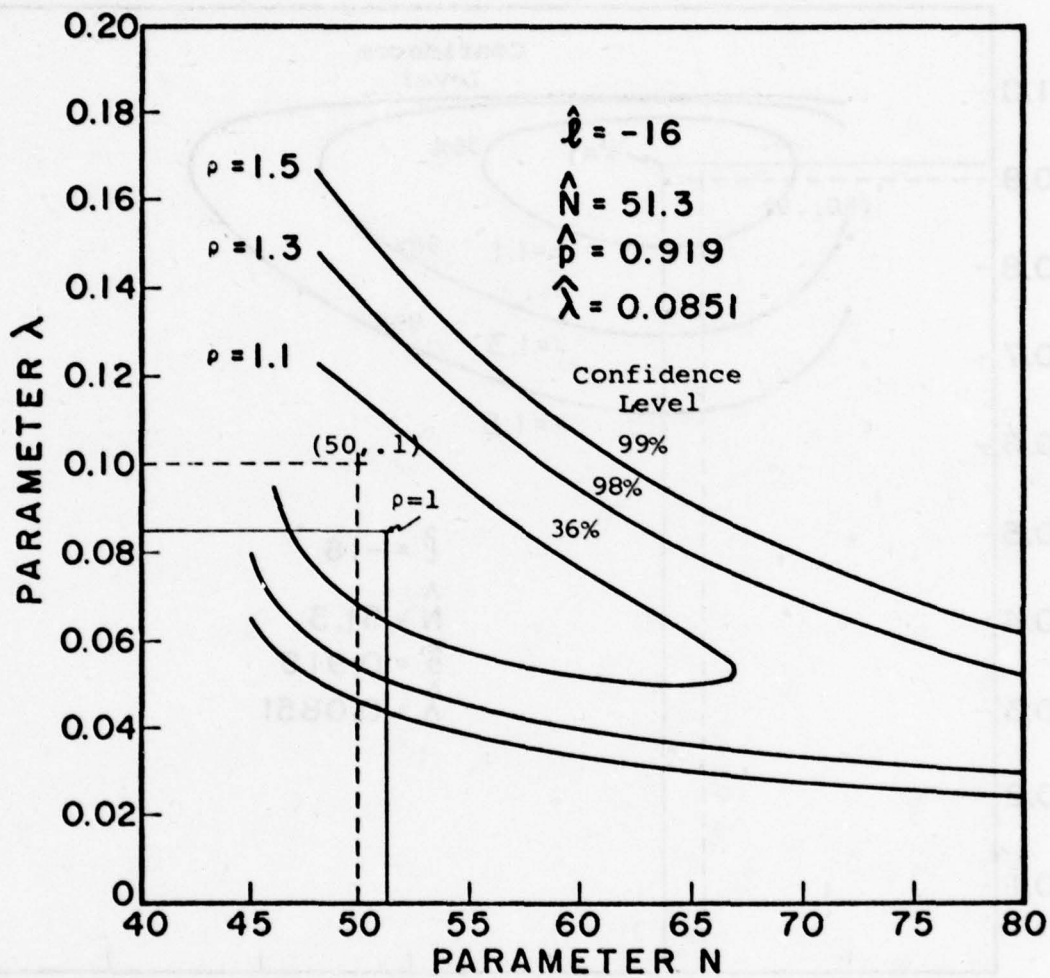


Figure 2.2 Likelihood Contours for N and λ when $p = \hat{p}$

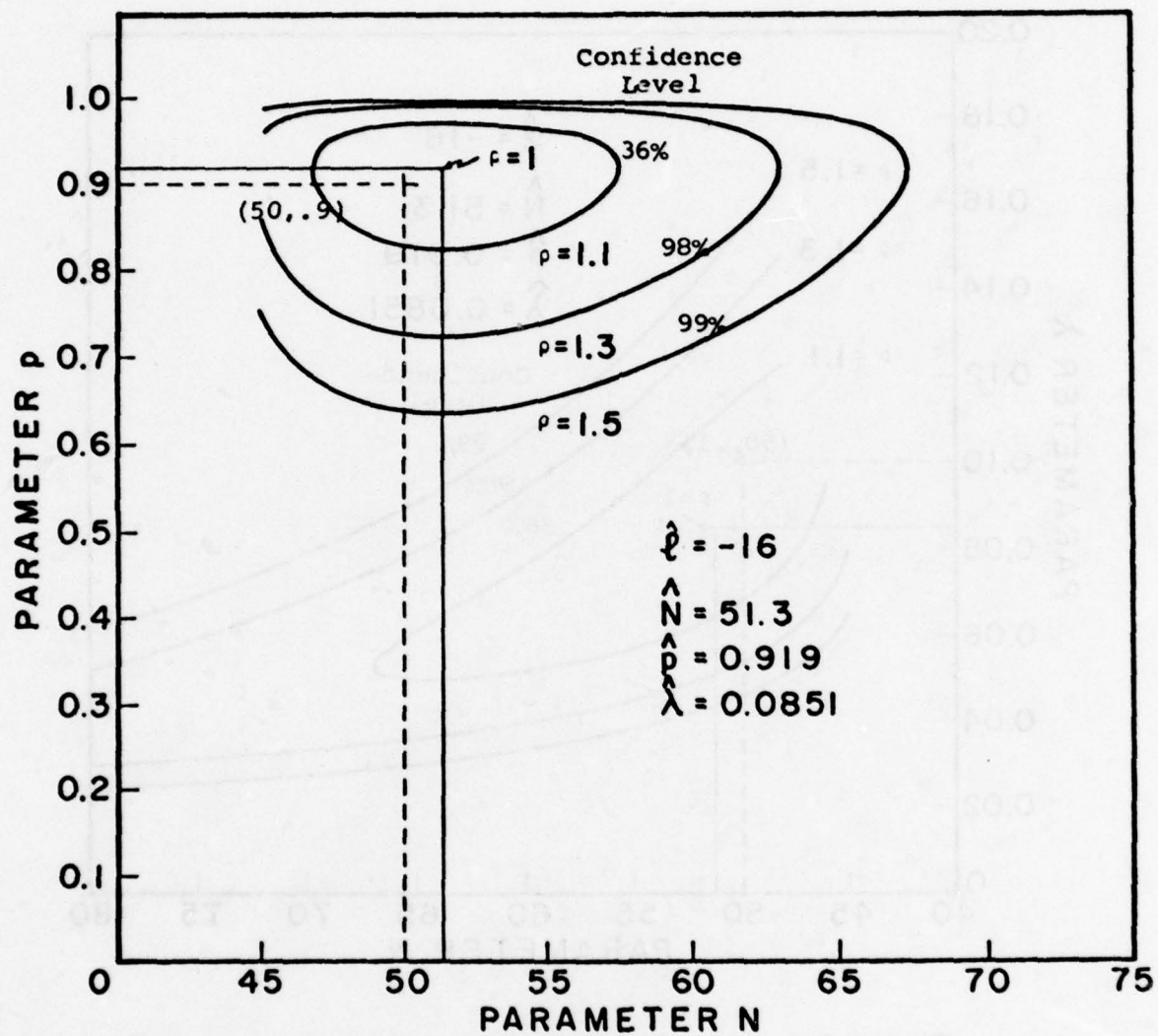


Figure 2.3 Likelihood Contours for N and p when $\lambda = \hat{\lambda}$

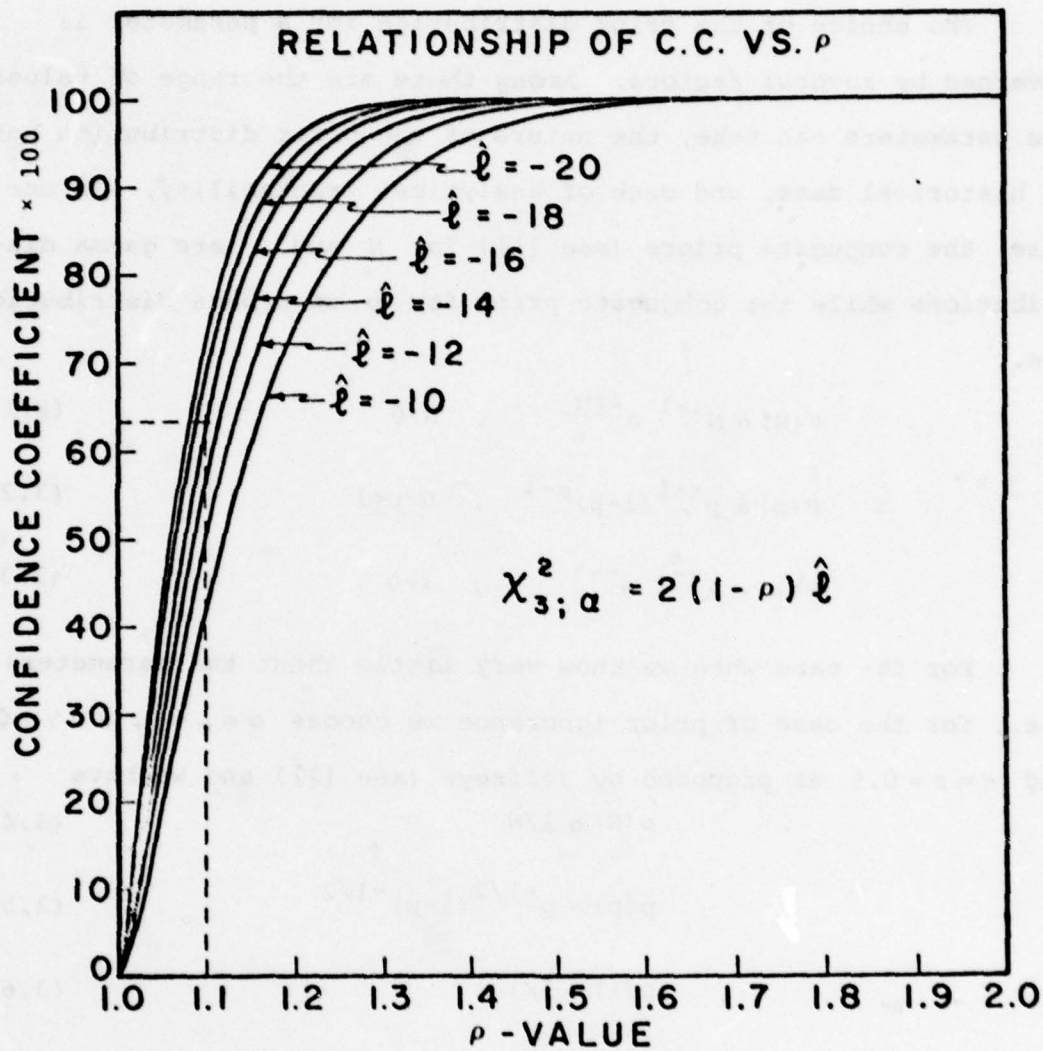


Figure 2.4 Relationship Between the Confidence Coefficient and the Constant ρ

3. BAYESIAN INFERENCE

In this section we use a Bayesian approach for obtaining posterior point estimates and the highest posterior density (HPD) region for parameters N , p and λ .

3.1 Prior Distributions

The choice of the prior distribution for a parameter is governed by several factors. Among these are the range of values the parameters can take, the nature of the prior distribution based on historical data, and case of analytical tractability. In our case, the conjugate priors (see [1]) for N and λ are gamma distributions while the conjugate prior for p is a beta distribution, i.e.,

$$P(N) \propto N^{\alpha-1} e^{-\beta N}, \quad N > 0 \quad (3.1)$$

$$P(p) \propto p^{\pi-1} (1-p)^{\rho-1}, \quad 0 \leq p \leq 1 \quad (3.2)$$

$$P(\lambda) \propto \lambda^{\mu-1} e^{-\gamma \lambda}, \quad \lambda > 0. \quad (3.3)$$

For the case when we know very little about the parameters i.e., for the case of prior ignorance we choose $\alpha = \mu = 0$, $\beta = \gamma = 0$ and $\pi = \rho = 0.5$ as proposed by Jeffreys (see [1]) and we have

$$p(N) \propto 1/N \quad (3.4)$$

$$p(p) \propto p^{-1/2} (1-p)^{-1/2} \quad (3.5)$$

$$p(\lambda) \propto 1/\lambda. \quad (3.6)$$

These are called the non-informative prior distributions.

We also assume the independence of prior information about N , p and λ , i.e.

$$p(N, p, \lambda) = p(N)p(p)p(\lambda). \quad (3.7)$$

3.2 Joint Posterior Distribution

By applying Bayes theorem we obtain the joint posterior distribution of N , p and λ for given priors and the data i.e.

$$p(N, p, \lambda | \underline{t}, \underline{y}) \propto p(N, p, \lambda) L(N, p, \lambda | \underline{t}, \underline{y}) \quad (3.8)$$

where the likelihood function $L(N, p, \lambda | \underline{t}, \underline{y})$ for given $\underline{t}$ and $\underline{y}$ is given by equation (2.5).

Let $\hat{N}$, $\hat{p}$ and $\hat{\lambda}$ be the Bayesian point estimates for N , p and λ , respectively. That is, the point $(\hat{N}, \hat{p}, \hat{\lambda})$ is the mode of the joint posterior distribution $p(N, p, \lambda | \underline{t}, \underline{y})$. In other words, $p(N, p, \lambda | \underline{t}, \underline{y})$ attains its maximum at $(\hat{N}, \hat{p}, \hat{\lambda})$. Therefore, $\hat{N}$, $\hat{p}$ and $\hat{\lambda}$ must satisfy

$$\frac{\partial}{\partial N} \log p(N, p, \lambda | \underline{t}, \underline{y}) = 0 \quad (3.9)$$

$$\frac{\partial}{\partial p} \log p(N, p, \lambda | \underline{t}, \underline{y}) = 0 \quad (3.10)$$

$$\frac{\partial}{\partial \lambda} \log p(N, p, \lambda | \underline{t}, \underline{y}) = 0 \quad (3.11)$$

where

$$\begin{aligned} \log p(N, p, \lambda | \underline{t}, \underline{y}) \propto & n \log \lambda - \sum_{i=1}^n \{N - p(i-1)\} t_i \\ & + \sum_{i=1}^n y_i \log(1-p) + \sum_{i=1}^n (1-y_i) \log(N - (i-1)) \\ & + (\alpha-1) \log N - \beta N \\ & + (\mu-1) \log \lambda - \gamma \lambda \\ & + (\pi-1) \log p + (\rho-1) \log(1-p) . \end{aligned} \quad (3.12)$$

Then we get

$$-\lambda \sum_{i=1}^n t_i + \sum_{i=1}^n \frac{1-y_i}{N-(i-1)} + \frac{\alpha-1}{N} - \beta = 0 \quad (3.13)$$

$$\lambda \sum_{i=1}^n (i-1)t_i - \sum y_i / (1-p) + \frac{\pi-1}{p} - \frac{\rho-1}{1-p} = 0 \quad (3.14)$$

$$n/\lambda - \sum_{i=1}^n (N-p(i-1))t_i + \frac{\mu-1}{\lambda} - \gamma = 0 \quad (3.15)$$

Simultaneous non-linear equations (3.13), (3.14) and (3.15) can be solved by numerical methods discussed in Section 2.1.

3.3 H.P.D. Regions

It is useful to obtain the Bayesian confidence region or H.P.D. region which gives the probability content of a contour for the joint posterior distribution, $p(N, p, \lambda | \underline{t}, \underline{y})$. As an approximation, we may use the fact that for large samples $p(N, p, \lambda | \underline{t}, \underline{y})$ tends to normality (see Box and Tiao [1]). Therefore,

$$-2 \log \frac{p(N, p, \lambda | \underline{t}, \underline{y})}{p(\hat{N}, \hat{p}, \hat{\lambda} | \underline{t}, \underline{y})} \sim \chi^2_3 \quad (3.16)$$

It follows that the contour defined by

$$\log p(N, p, \lambda | \underline{t}, \underline{y}) = \log p(\hat{N}, \hat{p}, \hat{\lambda} | \underline{t}, \underline{y}) - \frac{1}{2} \chi^2_{3; \alpha} \quad (3.17)$$

encloses a region whose probability content is approximately $(1-\alpha)$. Then the $100(1-\alpha)\%$ H.P.D. region is given by

$$f(N, p, \lambda) = C \quad (3.18)$$

where

$$\begin{aligned}
f(N, p, \lambda) = & n \log \lambda - \sum_{i=1}^n [N - p(i-1)] t_i \\
& + \sum_{i=1}^n y_i \log(1-p) + \sum_{i=1}^n (1-y_i) \log[N - (i-1)] \\
& + (\alpha-1) \log N - \beta N \\
& + (\mu-1) \log \lambda - \gamma \lambda \\
& + (\pi-1) \log p + (\rho-1) \log(1-p)
\end{aligned} \tag{3.19}$$

and

$$c = f(\hat{N}, \hat{p}, \hat{\lambda}) - \frac{1}{2} x_{3, \alpha}^2 . \tag{3.20}$$

The contour defined by (3.18) can be evaluated by numerical methods as discussed in Section 2.1.

3.4 Numerical Example

To illustrate the computations for the various quantities given in Sections 3.1, 3.2 and 3.3 we use the simulated data of Table 2.1. Using the non-informative priors given in equations (3.4), (3.5) and (3.6), the Bayesian point estimates of N , p , λ are obtained by solving equations (3.13), (3.14) and (3.15) and are

$$\hat{N} = 51.43$$

$$\hat{p} = 0.927$$

$$\hat{\lambda} = 0.0836 .$$

The Bayesian H.P.D. region for N , p and λ for this data set is obtained from equation (3.18). Taking $\alpha = .10$, the 90% H.P.D. region is shown in Figure 3.1.

The 50%, 75% and 90% Bayesian regions for p and λ when $N = \hat{N}$ are given in Figure 3.2. Similar regions for N and λ ($p = \hat{p}$) and for N and p ($\lambda = \hat{\lambda}$) are given in Figures 3.3 and 3.4, respectively.

It is also useful to study the shapes of the posterior distributions of parameters N , p and λ . These are obtained by fixing the other two parameters at their Bayesian point estimates. Plots of such distributions are given in Figures 3.5, 3.6 and 3.7 for N , p and λ , respectively.

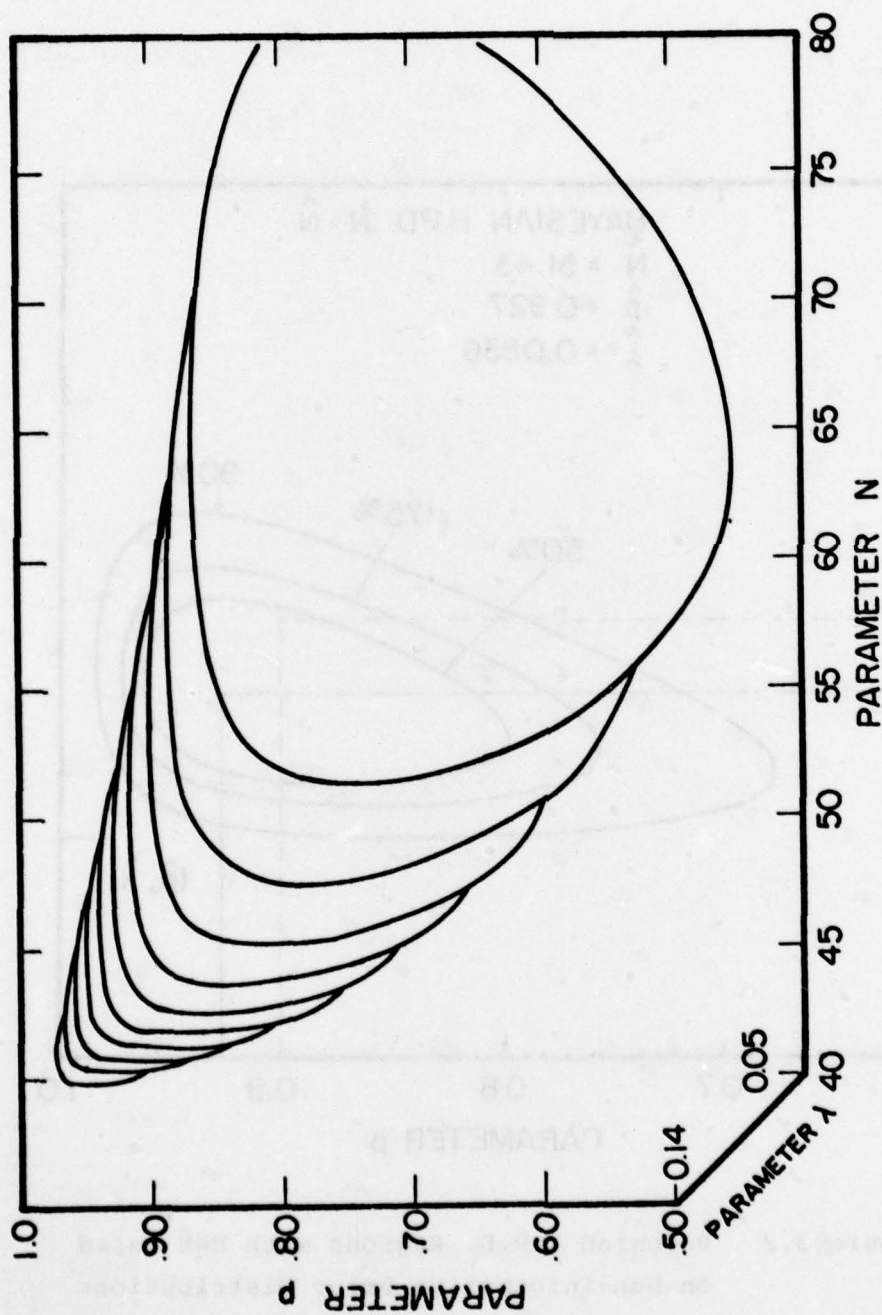


Figure 3.1 90% Bayesian H.P.D. Region for N , p and λ
for the Data in Table 2.1

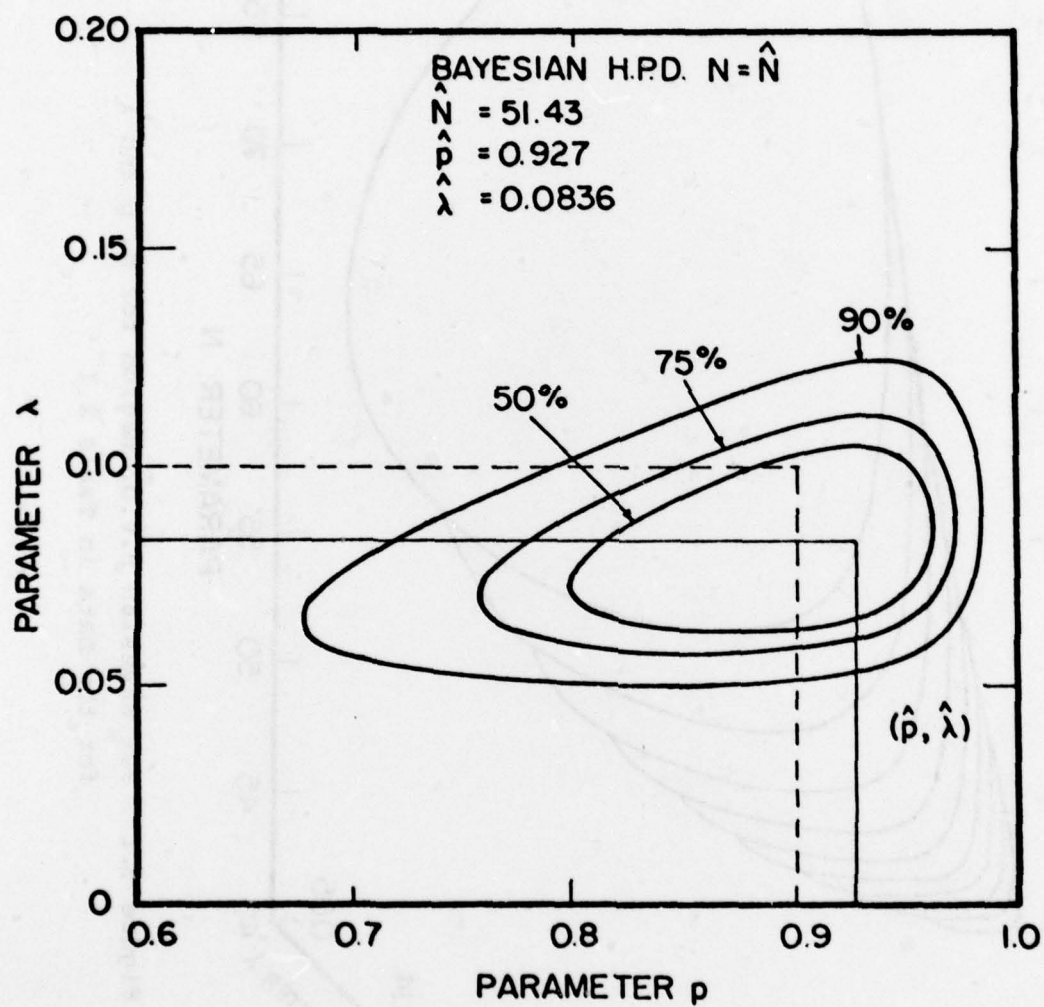


Figure 3.2 Bayesian H.P.D. Regions with $N = \hat{N}$ Based on Non-informative Prior Distributions

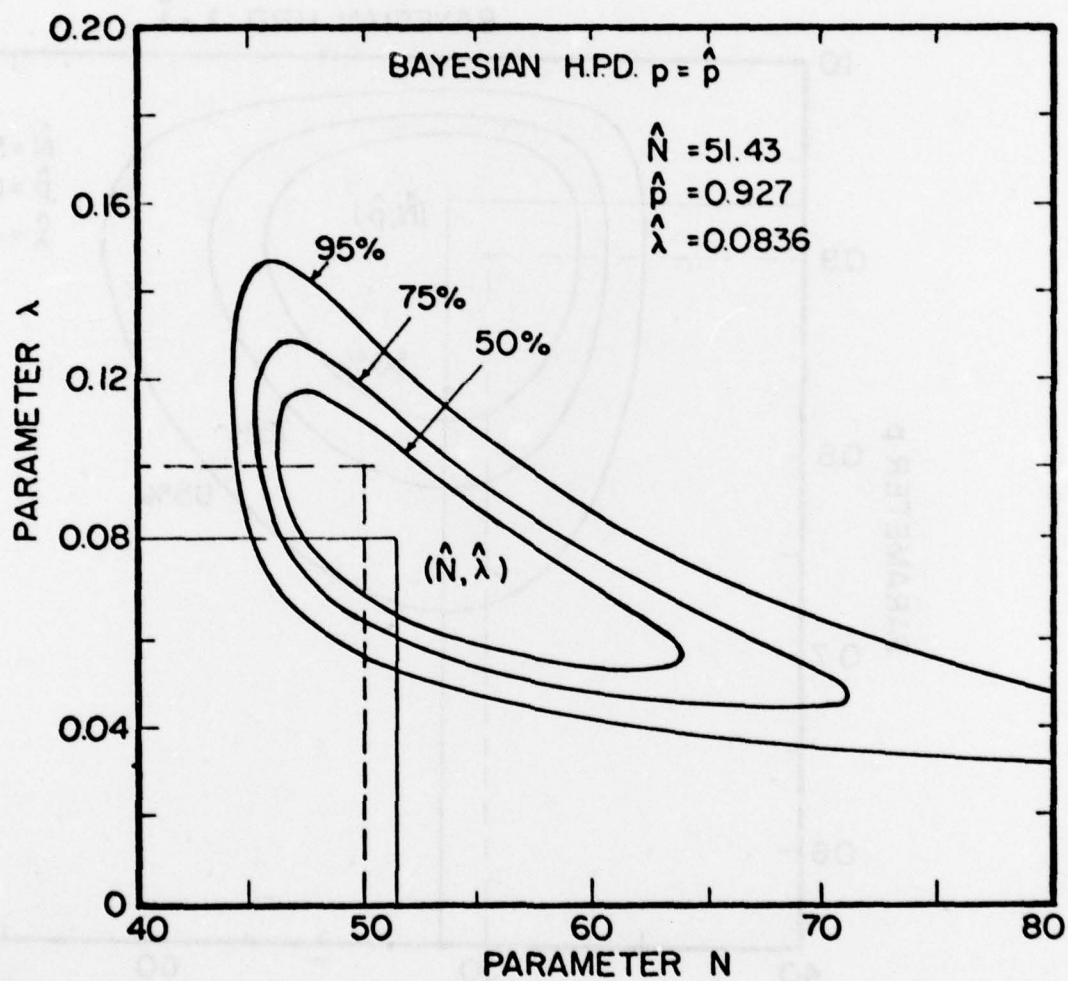


Figure 3.3 Bayesian H.P.D. Regions with $p = \hat{p}$ Based on Non-informative Prior Distributions

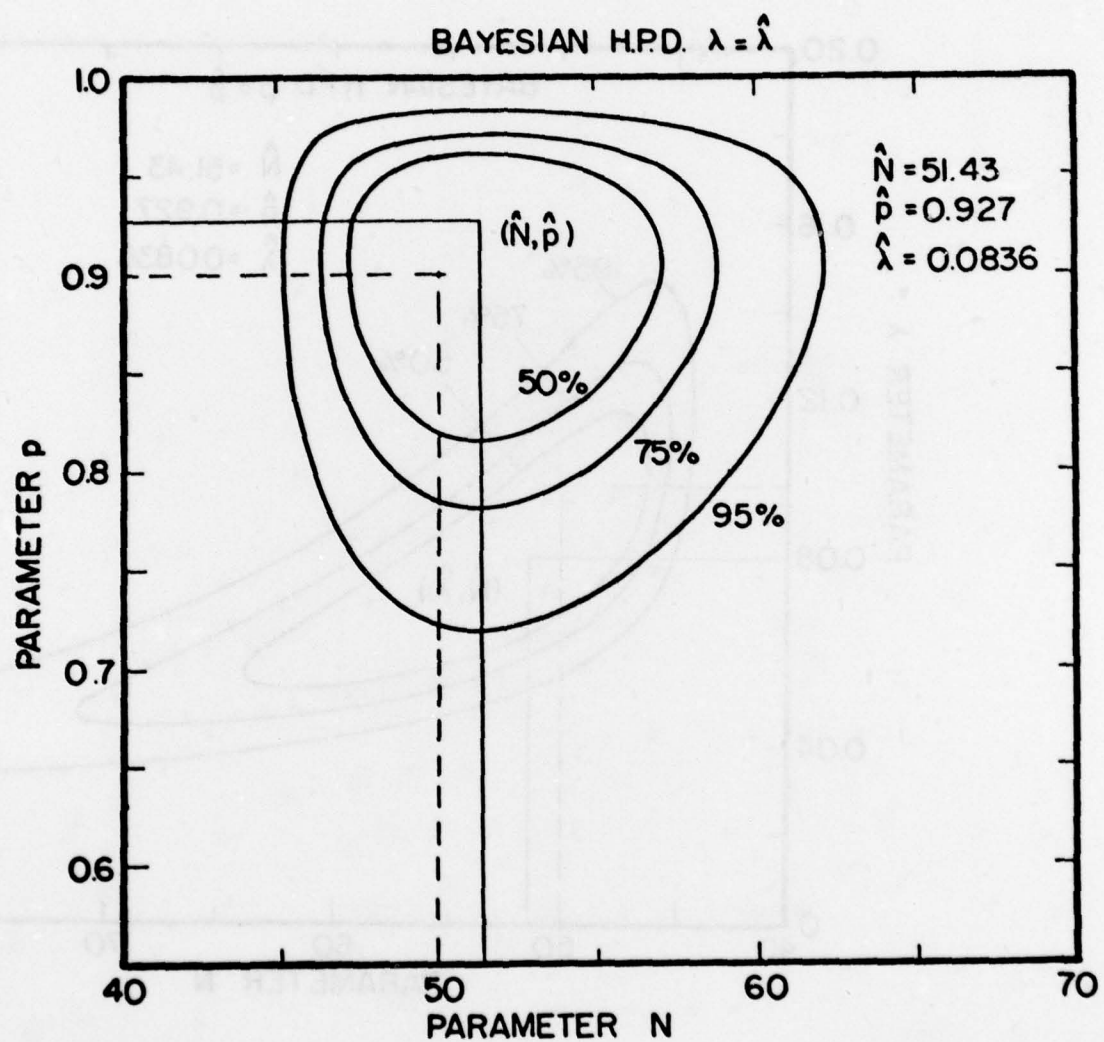


Figure 3.4 Bayesian H.P.D. Regions with $\lambda = \hat{\lambda}$ Based on Non-informative Prior Distributions

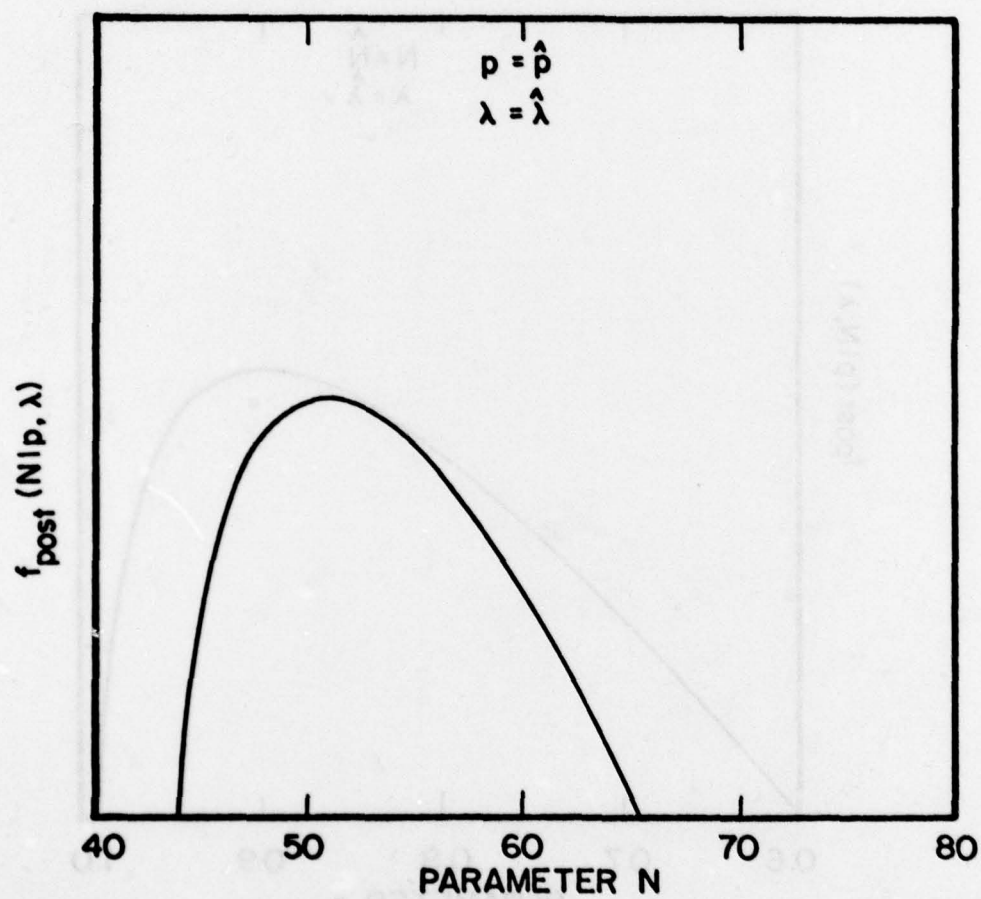


Figure 3.5 Posterior Distribution of Parameter N when $p=\hat{p}$ and $\lambda=\hat{\lambda}$

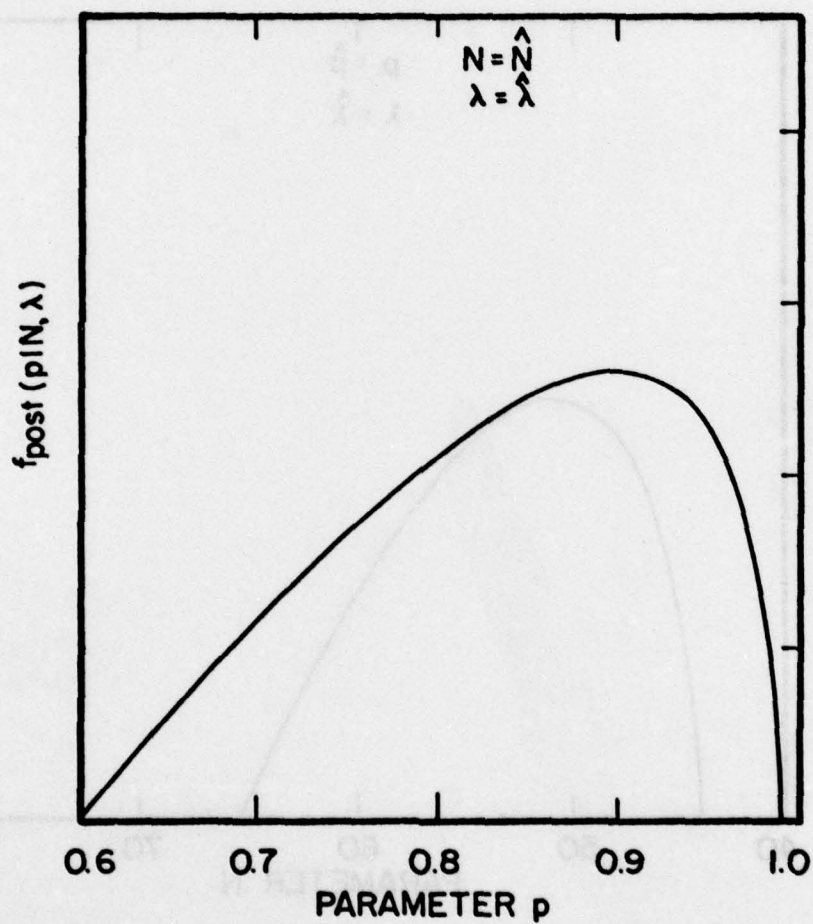


Figure 3.6 Posterior Distribution of Parameter p when $N = \hat{N}$ and $\lambda = \hat{\lambda}$

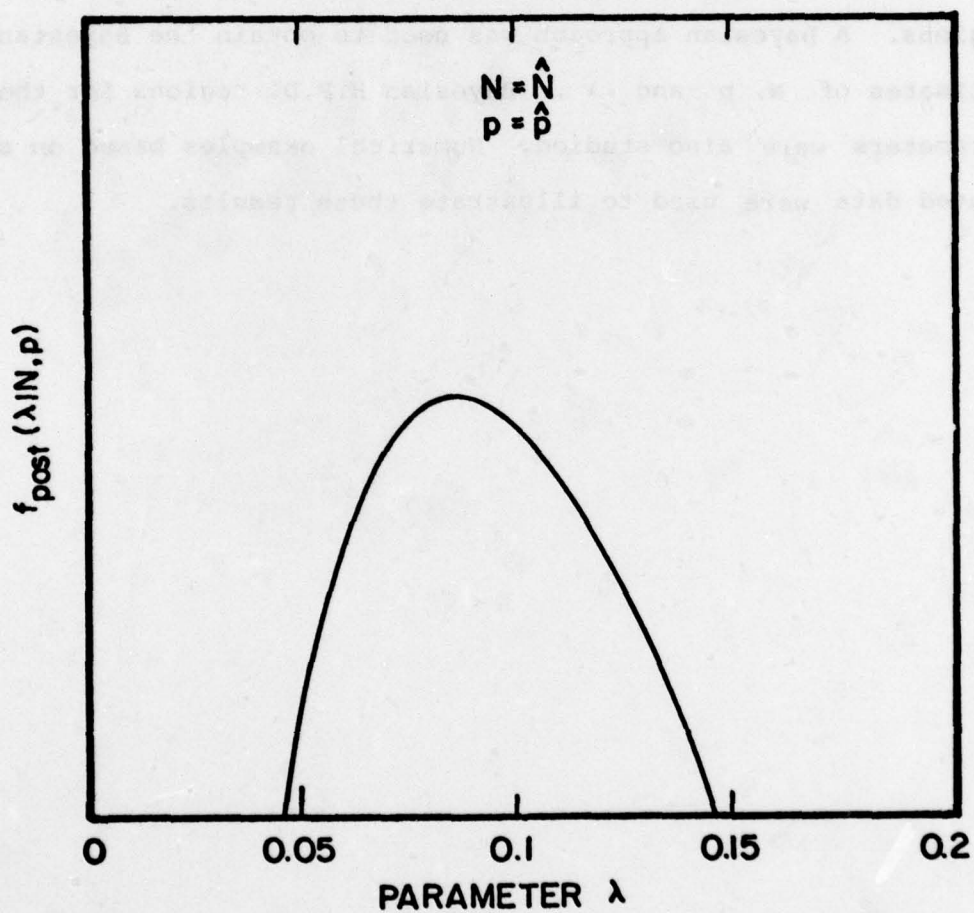
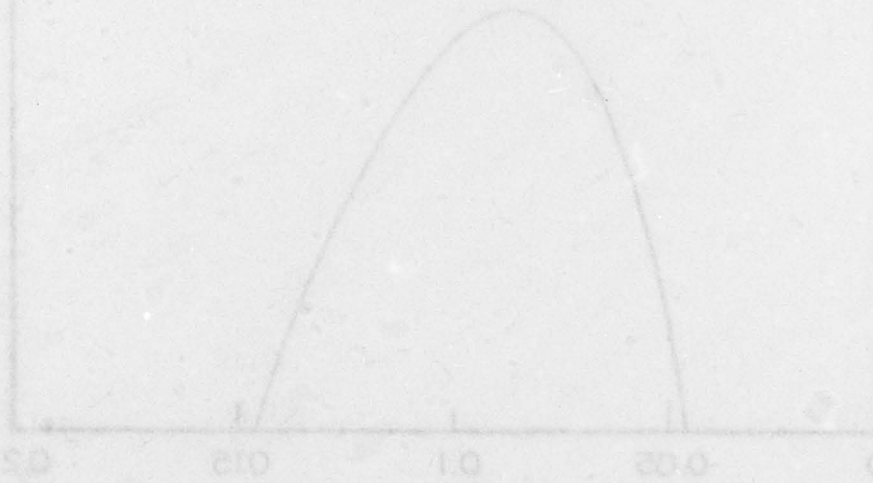


Figure 3.7 Posterior Distribution of Parameter λ when $N = \hat{N}$ and $p = \hat{p}$

4. CONCLUDING REMARKS

In this report we presented two methods for statistical inference of the parameters N , p and λ for the imperfect debugging model. Using the method of maximum likelihood, expressions were derived for the mle's, the likelihood contours and the confidence regions. A Bayesian approach was used to obtain the Bayesian point estimates of N , p and λ . Bayesian H.P.D. regions for these parameters were also studied. Numerical examples based on simulated data were used to illustrate these results.



5. REFERENCES

- [1] Box, G. E. P. and Tiao, G. C. (1973), Bayesian Inference in Statistical Analysis, Addison-Wesley.
- [2] Goel, A. L. and Okumoto, K. (1978), "An Imperfect Debugging Model for Reliability and Other Quantitative Measures of Software Systems," Technical Report No. 78-1, Department of IE & OR, Syracuse University.
- [3] Roussas, G. G. (1973), A First Course in Mathematical Statistics, Addison-Wesley.
- [4] Taha, H. A. (1976), Operations Research: An Introduction, 2nd Ed., MacMillan.

APPENDIX A SIMULATION OF DATA (t, y)

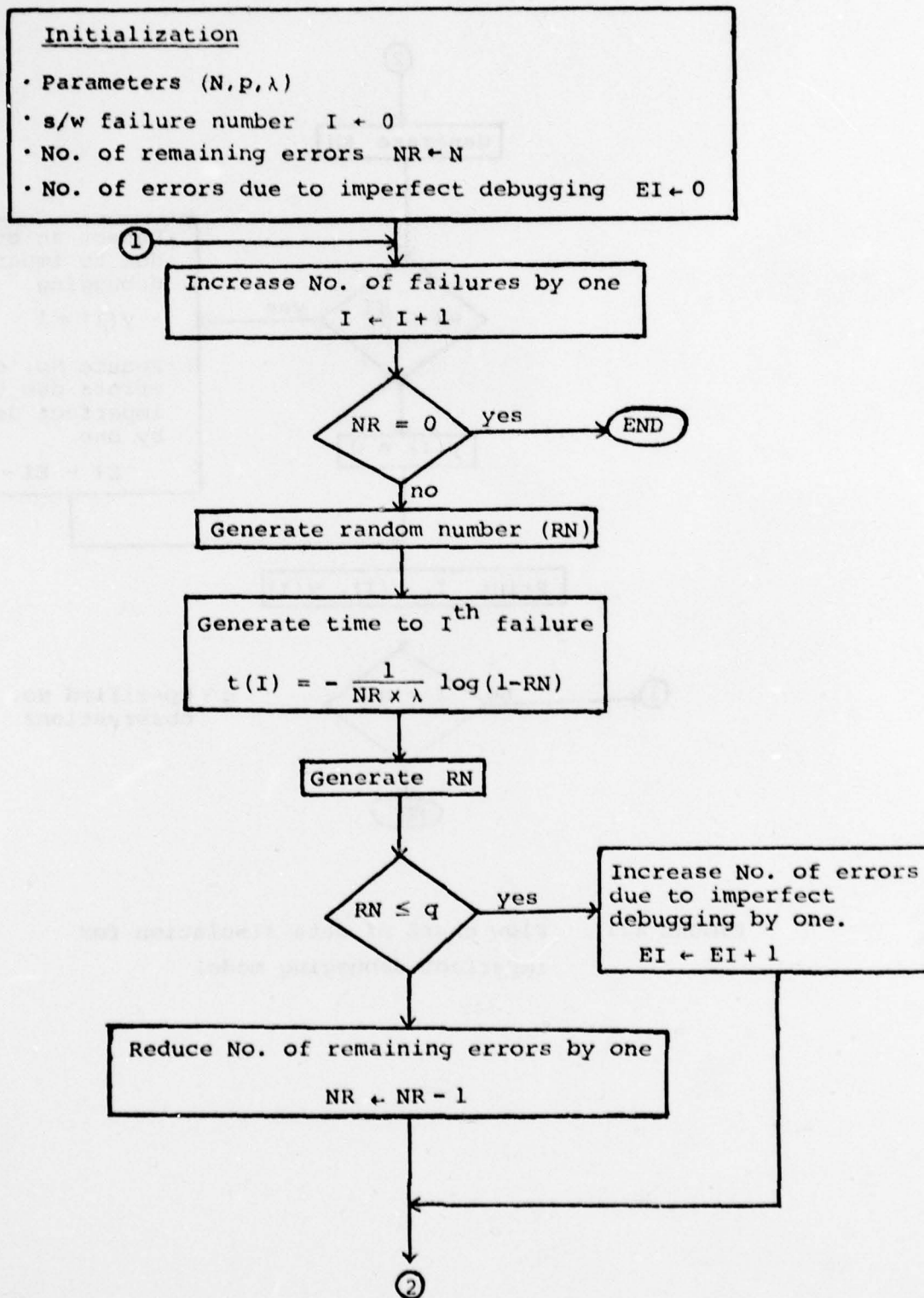
In this Appendix we describe the procedure used for simulating the data on times between software failures and the categories of errors. Recall that t_i denotes the time between the $(i-1)$ st and i th software failures. Also, assuming that a software error can be identified as being the one due to imperfect debugging, whenever it occurs, we have

$$y_i = \begin{cases} 1, & \text{if } i\text{th failure is caused by the error due to} \\ & \text{imperfect debugging} \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, t_i and y_i are the data needed for statistical inference of parameters N , p and λ in the imperfect debugging model.

A flow chart for simulating these data is given in Figure A.1. First we initialize the parameters N , λ , p , I (software failure number), NR (number of remaining errors) and EI (number of errors due to imperfect debugging). Then a random number RN which is uniformly distributed over $(0,1)$ is generated. Now, from equation (2.1), the random variable T_i has an exponential distribution with parameter $NR \cdot \lambda$; i.e.

$$F_{T_i}(t_i) = 1 - e^{-NR \cdot \lambda \cdot t_i}. \quad (A.1)$$



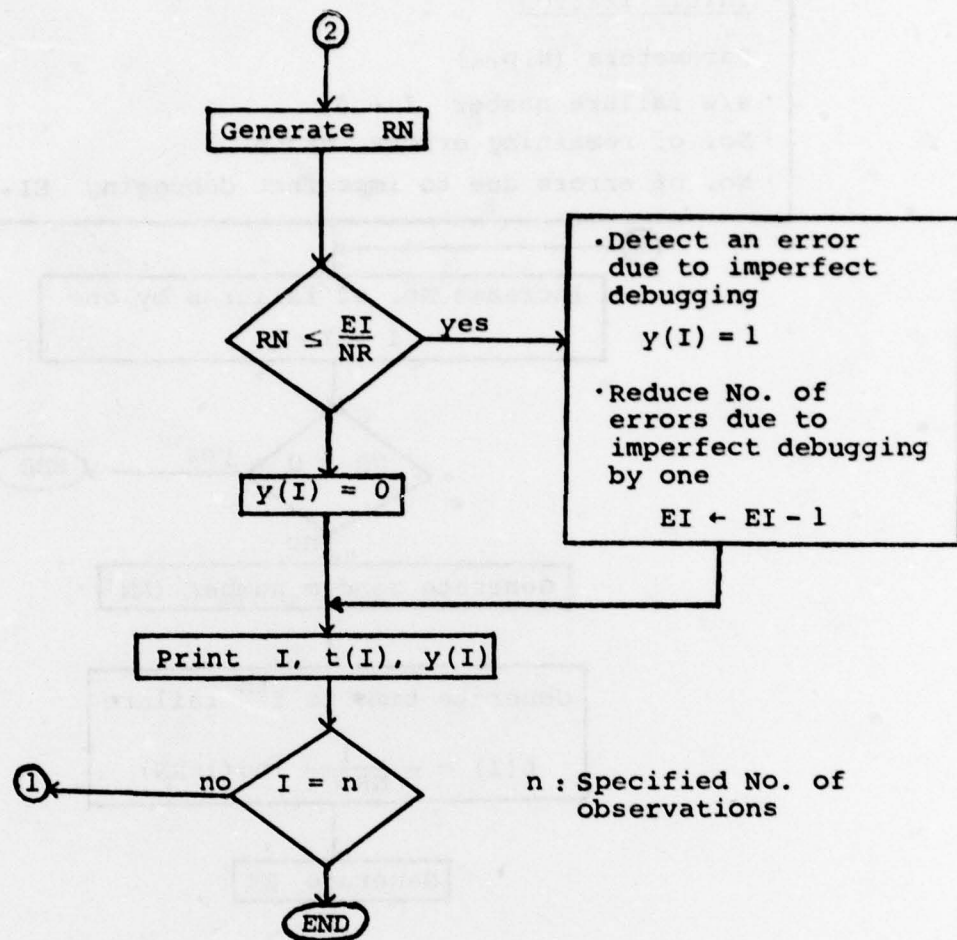


FIGURE A.1 Flow chart of data simulation for imperfect debugging model

For some value RN

$$1 - e^{-NR \cdot \lambda \cdot t_i} = RN \quad (A.2)$$

and hence the simulated value of t_i is given by

$$t_i = - \frac{1}{NR \cdot \lambda} \cdot \log(1-RN) . \quad (A.3)$$

Next, we generate a new random number RN. If this new number $RN \leq q$, the probability of imperfect debugging, the quantity EI is incremented by 1 and the number of remaining errors remains unchanged. If $RN > q$, the number of remaining errors NR is decreased by 1. An error which occurs next is selected randomly. For given EI and NR the probability that an error due to imperfect debugging is detected is EI/NR . Hence, if for a still new random number RN, $RN \leq EI/NR$, then we set $y_i = 1$ and decrease EI by 1. Otherwise, $y_i = 0$. After repeating this procedure n times, we obtain the simulated data set $(\underline{t}, \underline{y})$ where $\underline{t} = (t_1, t_2, \dots, t_n)$ and $\underline{y} = (y_1, y_2, \dots, y_n)$. Table 2.1 shows a data set simulated by this procedure, where $N = 50$, $p = 0.9$, $\lambda = 0.1$ and $n = 45$.