LANGLEY
GRANT
IN-64-CR

DEPARTMENT OF MATHEMATICS & STATISTICS
COLLEGE OF SCIENCES
OLD DOMINION UNIVERSITY
NORFOLK, VIRGINIA 23529

ESTIMATING RESIDUAL FAULT HITTING RATES BY RECAPTURE SAMPLING

By

Larry Lee, Principal Investigator

and

Rajan Gupta

Final Report
For the period ending December 31, 1988

Prepared for the
National Aeronautics and Space Administration
Langley Research Center
Hampton, Virginia

Under
Research Grant NAG-1-835
Dr. Dave E. Eckhardt, Jr. Technical Monitor
ISD-Systems Architecture Branch

Submitted by the
Old Dominion University Research Foundation
P.O. Box 6369
Norfolk, Virginia 23508

December 1988

Estimating Residual Fault Hitting Rates by Recapture Sampling

Rajan Gupta

Larry Lee*

Department of Mathematics and Statistics

Old Dominion University

Norfolk, VA. 23529-0077

SUMMARY

For the recapture debugging design introduced by Nayak (1988) we consider the problem of estimating the hitting rates of the faults remaining in a system. In the context of a conditional likelihood, moment estimators are derived and are shown to be asymptotically normal and fully efficient. Fixed sample properties of the moment estimators are compared, through simulation, with those of the conditional maximum likelihood estimators. Properties of the conditional model are investigated such as the asymptotic distribution of linear functions of the fault hitting frequencies and a representation of the full data vector in terms of a sequence of independent random vectors. It is assumed that the residual hitting rates follow a log linear rate model and that the testing process is truncated when

*This research was supported by the National Aeronautics and Space Administration, under contract NAG-1-835.

the gaps between the detection of new errors exceed a fixed amount of time.

Key words : Recapture Sampling; Software Reliability; Fault Hitting Rates; Asymptotic efficiency; Conditional Likelihood; Average Information; Interval Censored Sampling.

1. Introduction

Recapture sampling was introduced by Nayak (1988) as a way to get extra information for estimating the number of faults remaining in a system. By placing counters in the software we observe, in addition to the usual sequence of failure (i.e., error detection) times, the hitting frequencies of detected faults. Nayak's (1988) discussion concerns the Jelinski-Moranda (1972) model and procedures useful for estimating the number of remaining faults.

The present paper deals with the problem of estimating the residual fault hitting rates under a more flexible form of the model. Let λ_i be the hitting rate of the faults remaining in a system when $i - 1$ faults have been corrected. Since eliminating faults improves reliability, we have $\lambda_1 > \lambda_2 >. . .$ and the effects of correcting the first, second, etc., detected faults can be defined by $\xi_i = \lambda_i - \lambda_{i+1}$. To model the failure times T_1, T_2, ..., we assume that the failure gaps $Y_i = T_i - T_{i-1}$, $(T_0 \equiv 0)$ are independent and exponentially distributed with rate parameters λ_i; note that Y_i can be interpreted as the smallest of the times of encountering the remaining faults. As for Nayak's (1988) model, counts $\{M_i(t)\}$ of repeated error occurrences are assumed to be independent homogeneous Poisson processes with rate parameters ξ_i.

In attempting to get consistent estimators of the λ_i, the following difficulties arise: (i) the likelihood function based on the ordered failure times and hitting frequencies is not indexed by a measure of the amount of information and (ii) the number of parameters increases in proportion to the amount of information.

In view of (ii) we study estimation under the following model:

$$\lambda_i = \alpha e^{-\beta(i-1)}, \qquad \alpha, \ \beta > 0 \tag{1}$$

$$\xi_i = \alpha e^{-\beta(i-1)}(1 - e^{-\beta}) \quad i = 1, 2, ...$$

This model yields a tractable form of the likelihood function and is considerably more flexible than Jelinski-Moranda (1972) model. The latter model $\lambda_i = (\nu - i + 1)\phi, i = 1, 2, ..., \nu, \ \phi > 0$ assumes a common rate, $\xi_i = \phi$. Empirical evidence indicates, however, that faults may have different hitting rates (Nagel, Scholz, Skrivan, 1984). The effects of correcting faults will typically decrease since faults having the highest hitting rates are likely to be detected early during the debugging process.

The full likelihood function is presented in Section 2. This likelihood, however, is not indexed by a measure of the amount of information so we consider in Section 3, a conditional likelihood obtained by conditioning on the observed number of detected faults. In this case the sufficient statistic can be represented as a sum of independent nonidentically distributed random vectors. Using this representation, we show in Section 4 that certain moment estimators are asymptotically normal and fully efficient. In Section 4 we also consider estimating on upper bound for λ_{r+1} where r is the number of faults eventually corrected. Fixed sample properties of the moment estimators and the conditional maximum likelihood estimators are evaluated, through simulation, in Section 5.

2. NOTATION AND OTHER PRELIMINIARIES

Suppose a system is tested until the failure gaps first exceed a fixed amount of time s. We observe $T_1, T_2, ..., T_R$ where R , the number of faults eventually detected, is determined by truncating $Y_i = T_i - T_{i-1}, i = 1, 2, \ldots$ over the interval $(0, s)$. That is, data is obtained through an interval censoring model in which we observe the realization $R = r$ providing $Y_i \leq s, i = 1, 2, ..., r$ and $Y_{r+1} > s$. The total test time $\tau = \sum_1^R Y_i + s$ is then random while for Nayak's (1988) censoring model, τ is a fixed time of testing.

Since $Y_1, Y_2, ...,$ are assumed to be independent exponential random variables, the joint density function of $(R, Y_1, Y_2, \ldots, Y_R)$ is

$$f(y_1, y_2, \ldots, y_r; r) = \exp(-\lambda_{r+1}s) \prod_{i=1}^{r} \lambda_i \exp(-\lambda_i y_i), \quad 0 < y_i < s, i = 1, 2, \ldots, r \tag{2}$$

Explicit formulas for the case $r = 0$, will be omitted since if $r = 0$, there is little information available for inference.

Recall that counts $\{M_i(t)\}$ of repeated error occurrences are assumed to be independent Poisson processes with rate parameters ξ_i. For the fault detected at time t_i let M_i be the number of times this fault is accessed during the interval $(t_i, \tau]$. Given $t_1, t_2, ..., t_r$ the vector $(M_1, M_2, ..., M_r)$ consists of independent Poisson random variables with means $\xi_i(\tau - t_i), i = 1, 2, ..., r$.

The joint density of $(R, Y_1, Y_2, ..., Y_R,\ M_1, M_2, ..., M_R)$ is

$$exp(-\lambda_{r+1}s) \prod_{i=1}^{r} \lambda_i \exp(-\lambda_i y_i) \prod_{i=1}^{r} [\xi_i(\sum_{j=i+1}^{r} y_j + s)]^{m_i} exp[-\xi_i(\sum_{j=i+1}^{r} y_j + s)]/m_i!$$

$$= \exp(-\lambda_1 \tau) \prod_{i=1}^{r} \lambda_i [\xi_i(\sum_{j=i+1}^{r} y_j + s)]^{m_i}/m_i! \tag{3}$$

Under (1) the likelihood function given by (3) is maximized by

$$\hat{\alpha} = (r + \sum_{1}^{r} m_i)/\tau$$

$$\hat{\beta} = ln\{1 + \sum_{1}^{r} m_i/[\sum_{1}^{r}(i-1)m_i + r(r-1)/2]\}$$

However, since this likelihood is not indexed by a measure of the amount of information, the asymptotic properties of these estimators cannot be deduced from the form of the likelihood.

3. A Conditional Likelihood

Under the censoring model of Section 2, the conditional density of $Y_1, Y_2, \ldots, Y_r$, given $R = r$, is

$$\prod_{i=1}^{r} \lambda_i \exp(-\lambda_i y_i)[1 - \exp(-\lambda_i s)]^{-1}, \quad 0 < y_i < s; \ i = 1, 2, \ldots, r \qquad (4)$$

Consider the sequence defined by

$$Z_1 = (Y_1), \quad Z_k = (Y_k, M_{1k}, M_{2k}, \ldots, M_{(k-1)k}) \quad (k = 2, 3, ..., r+1) \qquad (5)$$

where $M_{ik}, i < k$, is the number of repeated occurrences of error i during the interval $(T_{k-1}, T_k]$ and, for convenience, we let $Y_{r+1} = s$. The last interval $(T_r, \tau]$ has fixed length s while $(T_{k-1}, T_k]$, $k \leq r$, has random length Y_k. Our earlier assumption that $\{M_i(t)\}$ are independent homogeneous Poisson processes implies that $\{M_{ik}\}$, $i < k, k = 2, 3, ..., r+1$ are conditionally, given $Y_1, Y_2, ..., Y_R$ and R, independent Poisson random variables with means $\xi_i y_k$.

Since counts of events occurring in different intervals are independent and the interval lengths are also independent, it follows that $Z_1, Z_2, ..., Z_{r+1}$ are conditionally, given $R = r$, independent with densities

$$\begin{aligned} & g_k(y_k; m_{1k}, m_{2k}, \ldots, m_{(k-1)k}) \\ & = \lambda_k e^{-\lambda_k y_k}(1 - e^{-\lambda_k s})^{-1} \prod_{i=1}^{k-1} (\xi_i y_k)^{m_{ik}} e^{-\xi_i y_k}/m_{ik}! \quad (k = 1, 2, ..., r) \\ & = \prod_{i=1}^{r} (\xi_i s)^{m_{i(r+1)}} \, e^{-\xi_i s}/m_{i(r+1)}! \quad (k = r+1) \end{aligned} \tag{6}$$

Substituting our model for λ_i and ξ_i in (6) and simplifying by using $\sum_{i=1}^{k-1} \xi_i = \lambda_1 - \lambda_k$, gives the conditional log likelihood

$$l_k = C_k(\alpha, \beta) + ln(1 - e^{-\beta}) \sum_{i=1}^{k-1} m_{ik} + \ln \alpha \sum_{i=1}^{k-1} m_{ik} - \beta \sum_{i=1}^{k-1} (i-1) m_{ik} - \alpha y_k + C \tag{7}$$

Here C does not depend upon (α, β), and

$$\begin{aligned} C_k(\alpha, \beta) &= ln\alpha - \beta(k-1) - ln\{1 - exp[-\alpha s e^{-\beta(k-1)}]\} \quad (k = 1, 2, ..., r) \\ &= \alpha s e^{-\beta r} \qquad (k = r+1) \end{aligned}$$

In terms of λ_k, $C_k(\alpha, \beta) = ln[\lambda_k(1 - e^{-\lambda_k s})^{-1}]$, $k = 1, 2, ..., r$ and $C_{r+1}(\alpha, \beta) = \lambda_{r+1} s$. Let $V_k = (V_{1k}, V_{2k}, V_{3k})$, $k = 1, 2, ..., r+1$ be defined by

$$V_{1k} = Y_k, \quad V_{2k} = \sum_{i=1}^{k-1} M_{ik}, \quad V_{3k} = \sum_{i=1}^{k-1} (i-1) M_{ik} \tag{8}$$

where, as before, $Y_{r+1} = s$. In reference to l_k in (7) and the full conditional likelihood $l_c = \sum_1^{r+1} l_k$ note the following:

(i) $V_1, V_2, ..., V_{r+1}$ are independent.

(ii) V_k is a minimal sufficient statistic for l_k.

(iii) $S_r = \sum_1^{r+1} V_k$ is a sufficient statistic for the family defined by l_c.

(iv) $S_r = (S_{1r}, S_{2r}, S_{3r})$ is given by $S_{1r} = r, \quad S_{2r} = \sum_1^r M_i, \quad S_{3r} = \sum_1^r (i-1) M_i$

To study the asymptotic distribution of S_r, we shall need the following :

THEOREM 1. *Let V_k be defined as in (8) where $(Y_k, M_{1k}, M_{2k}, ..., M_{(k-1)k})$, $k = 1, 2, \ldots, r+1$ are independent with densities (7). Let the means, variances and covariances of the elements of V_k be denoted by $\mu_{ik} = E(V_{ik})\,; i = 1, 2, 3$ and $\sigma_{ijk} = Cov(V_{ik}, V_{jk})$; $i, j = 1, 2, 3$. Then*

$$\mu_{1k} = \lambda_k^{-1} - s(e^{\lambda_k s} - 1)^{-1}$$

$$\mu_{2k} = \mu_{1k}(\lambda_1 - \lambda_k)$$

$$\mu_{3k} = \mu_{1k} a_k$$

$$\sigma_{11k} = \lambda_k^{-2} - s^2 e^{\lambda_k s}(e^{\lambda_k s} - 1)^{-2}$$

$$\sigma_{12k} = \sigma_{11k}(\lambda_1 - \lambda_k)$$

$$\sigma_{13k} = \sigma_{11k} a_k$$

$$\sigma_{22k} = \mu_{1k}(\lambda_1 - \lambda_k) + \sigma_{11k}(\lambda_1 - \lambda_k)^2$$

$$\sigma_{23k} = \mu_{1k} a_k + \sigma_{11k}(\lambda_1 - \lambda_k) a_k$$

$$\sigma_{33k} = \mu_{1k} b_k + \sigma_{11k} a_k^2$$

$$b_k = \sum_{i=1}^{k-1} (i-1)^2 \xi_i$$

$$a_k = \sum_{i=1}^{k-1} (i-1) \xi_i$$

$$\sigma_{11(r+1)} = 0$$

$$\mu_{1(r+1)} = s$$

PROOF: The mean μ_{1k} and variance σ_{11k} of Y_k can be obtained directly from the fact that Y_k has an exponential distribution truncated over the interval (0,s). Since V_{2k} is conditionally, given Y_k, a Poisson random variable with mean $Y_k \sum_{i=1}^{k-1} \xi_i = Y_k(\lambda_1 - \lambda_k)$, and $Var(V_{2k}) = E[Var(V_{2k}|Y_k)] + Var[E(V_{2k}|Y_k)]$ we immediately have the expression given for σ_{22k}. Similar calculations can be made for V_{3k} by noting that V_{3k} is a linear sum of the conditionally (given Y_k) independent Poisson random variables $M_{ik}, i < k$. These being routine calculations, further details are omitted.

The Average Information Matrix:

The second and mixed partial derivatives of l_k are as follows :

$$\frac{\partial^2 l_k}{\partial \alpha^2} = -(1+\sum_{i=1}^{k-1} M_{ik})\alpha^{-2} + s^2\alpha^{-2}\lambda_k^2 e^{\lambda_k s}(e^{\lambda_k s}-1)^{-2}$$

$$\frac{\partial^2 l_k}{\partial \alpha \partial \beta} = s\alpha^{-1}(k-1)\lambda_k(e^{\lambda_k s}-1)^{-1} - s^2\alpha^{-1}(k-1)\lambda_k^2 e^{\lambda_k s}(e^{\lambda_k s}-1)^{-2}$$

$$\frac{\partial^2 l_k}{\partial \beta^2} = -s(k-1)^2\lambda_k(e^{\lambda_k s}-1)^{-1} + s^2(k-1)^2\lambda_k^2 e^{\lambda_k s}(e^{\lambda_k s}-1)^{-2} - e^{\beta}(e^{\beta}-1)^{-2}\sum_{i=1}^{k-1} M_{ik}$$

The Fisher Information Matrix $A_k = (a_{ijk})$, based on l_k, is given by

$$a_{11k} = -E(\frac{\partial^2 l_k}{\partial \alpha^2}) = \alpha^{-1}\mu_{1k} + \alpha^{-2}(\lambda_k^2\sigma_{11k} - \lambda_k\mu_{1k})$$

$$a_{12k} = -E(\frac{\partial^2 l_k}{\partial \alpha \partial \beta}) = -(k-1)\alpha^{-1}(\lambda_k^2\sigma_{11k} - \lambda_k\mu_{1k})$$

$$a_{22k} = -E(\frac{\partial^2 l_k}{\partial \beta^2}) = (k-1)^2(\lambda_k^2\sigma_{11k} - \lambda_k\mu_{1k}) + (\lambda_1 - \lambda_k)\mu_{1k}e^{\beta}(e^{\beta}-1)^{-2}$$

To simplify the calculation of these quantities, write the second derivative expressions in terms of the moments of Y_k. For example,

$$s^2 e^{\lambda_k s}(e^{\lambda_k s}-1)^{-2} = \lambda_k^{-2} - \sigma_{11k}$$

$$\frac{\partial^2 l_k}{\partial \alpha^2} = -\alpha^{-2}\sum_{i=1}^{k-1} M_{ik} - \alpha^{-2}\lambda_k^2\sigma_{11k}$$

Then a_{11k} is obtained by using first the moment of V_{2k} given in Theorem 1. Similar calculations yield the expressions given for a_{12k} and a_{22k} .

Since the distribution of Y_k converges (as $k \to \infty$) to a uniform distribution on the interval $(0, s)$ the moments of Y_k converge to the corresponding moments of the limiting uniform distribution (Serfling, 1980 p.14). We, therefore, have $\lim_{k\to\infty}\mu_{1k} = s/2$, $\lim_{k\to\infty}\sigma_{11k} = s^2/12$, and $\lim_{k\to\infty}(k-1)^{\gamma}\lambda_k = 0$ for $\gamma = 0, 1, 2$. Using these limits to get the limiting average information matrix $A = \lim(1/r)(A_1 + A_2 + \ldots + A_{r+1})$, we have $A = (a_{ij})$ where $a_{11} = s/2\alpha$, $a_{12} = 0$, $a_{22} = \alpha s e^{\beta}(e^{\beta}-1)^{-2}/2$

4. ESTIMATION

Let $\tilde{\alpha}$ and $\tilde{\beta}$ denote estimators of α and β defined by

$$\tilde{\alpha} = \sum_{i=1}^{r} M_i/r$$

$$\tilde{\beta} = ln[1 + \sum_{i=1}^{r} M_i / \sum_{i=1}^{r}(i-1)M_i]$$

Here we derive $\tilde{\alpha}$ and $\tilde{\beta}$ by equating $(1/r)\sum_1^r M_i$ and $(1/r)\sum_1^r (i-1)M_i$ to the appropriate mean vector elements of the asymptotic distribution of $(1/r)S_r$.

THEOREM 2. *Under the assumptions of Theorem 1, $(1/r)S_r$ has a limiting (as r tends to infinity) normal distribution with mean vector $\mu' = (\mu_1, \mu_2, \mu_3)$ and covariance matrix $(1/r)\sum$, $\sum = (\sigma_{ij})$, where*

$$\mu_1 = s/2 \qquad \mu_2 = \alpha s/2 \qquad \mu_3 = \alpha s(e^\beta - 1)^{-1}/2$$

$$\sigma_{11} = s^2/12 \quad \sigma_{12} = \alpha s^2/12 \quad \sigma_{13} = \alpha s^2(e^\beta - 1)^{-1}/12$$

$$\sigma_{22} = \alpha s/2 + \alpha^2 s^2/12$$

$$\sigma_{23} = \alpha(e^\beta - 1)^{-1}[s/2 + \alpha s^2/12]$$

$$\sigma_{33} = (\alpha s/2)[(e^\beta - 1)^{-1} + 2(e^\beta - 1)^{-2}] + (\alpha^2 s^2/12)(e^\beta - 1)^{-2}$$

PROOF: The elements of μ and Σ are given by $\mu_i = lim(1/r)\sum_{k=1}^{r+1}\mu_{ik}, i = 1,2,3$ and $\sigma_{ij} = lim \ (1/r)\sum_{k=1}^{r+1}\sigma_{ijk}$ as r tends to infinity, where the terms in these sums are the moments given in Theorem 1. Since μ_{ik} and σ_{ijk} converge to finite limits as k tends to infinity, we have $\mu_i = lim \ \mu_{ik}$ and $\sigma_{ij} = lim \ \sigma_{ijk}$. Thus the calculations are similar to those discussed at the end of Section 3. The remainder of the proof requires showing

(Serfling, 1980, p.30) that

$$lim(1/r)\sum_{k=1}^{r+1} h_{kr} = 0 \tag{9}$$

where

$$h_{kr} = E[U_k I(U_k > \epsilon^2 r)]$$
$$U_k = \sum_{i=1}^{3}(V_{ik} - \mu_{ik})^2 \tag{10}$$

and $I(.)$ is the indicator function. Since $h_{kr} \leq (\epsilon^2 r)^{-1}E(U_k^2)$, the limit in (9) can be established by examining the fourth central and product moments of the V_{ik}. Further details are given in the Appendix.

Note that $\alpha = g_1(\mu_1, \mu_2, \mu_3)$ and $\beta = g_2(\mu_1, \mu_2, \mu_3)$ where $g_1(z_1, z_2, z_3) = z_2/z_1$, $g_2(z_1, z_2, z_3) = ln(1 + z_2/z_3)$, and μ' is the mean vector of the asymptotic distribution of $(1/r)S_r$. Applying the δ-method gives the following.

COROLLARY 1. *Under the assumptions of Theorem 1, $\tilde{\alpha}$ and $\tilde{\beta}$ have a limiting $(r \rightarrow \infty)$ normal distribution with mean vector (α, β) and are asymptotically independent with variances $\sigma_{\tilde{\alpha}}{}^2 = 2\alpha/rs$, $\sigma_{\tilde{\beta}}{}^2 = 2(e^\beta - 1)^2 e^{-\beta}/r\alpha s$.*

COROLLARY 2. *Under the assumptions of Theorem 1, $\tilde{\lambda}_{q+1} = \tilde{\alpha}e^{-\tilde{\beta}q}$, for fixed $q < r+1$, has a limiting $(r \rightarrow \infty)$ normal distribution with mean λ_{q+1} and variance*
$\sigma_q^2 = (2\alpha/rs)e^{-2\beta q}[1 + 2q^2(e^\beta - 1)^2 e^{-\beta}]$.

By comparing the form of $(\tilde{\alpha}, \tilde{\beta})$ with that of the maximum likelihood estimators given in Section 2, it is evident that the latter are not consistent. Consistency, however, depends upon the choice of index for the likelihood function. That is, there may be other ways of conditioning (i.e., other families of conditional likelihoods) which imply the consistency of $\hat{\alpha}$ and $\hat{\beta}$.

Corollary 2, gives an asymptotic basis for estimating an upper bound on λ_{r+1} but does not justify directly estimating λ_{r+1}. The hitting frequencies at the upper end of the vector $(M_1, M_2, ..., M_r)$ will tend to be small (often zero). In such cases, for a suitably chosen index $q < r+1$, $\lambda_{q+1} - \lambda_{r+1} = \sum_{i=q+1}^{r} \xi_i$ will tend to be small so that λ_{q+1} may be a close upper bound on λ_{r+1}.

5. Efficiency and Bias

Since ${\sigma_{\tilde{\alpha}}}^2 = (ra_{11})^{-1}$ and ${\sigma_{\tilde{\beta}}}^2 = (ra_{22})^{-1}$ where a_{11} and a_{22} are given at the end of Section 3, it follows that $\tilde{\alpha}$ and $\tilde{\beta}$ are asymptotically fully efficient.

To study the fixed sample properties of $\tilde{\alpha}$ and $\tilde{\beta}$, we simulated their values under the conditional model defined in (6). This was done by generating 1,000 replicates of $Z = (Z_1, Z_2, ..., Z_{r+1})$ for the values of r shown in Table 1. In addition the conditional maximum likelihood estimates $\hat{\alpha}_c$ and $\hat{\beta}_c$ were calculated for each replicate by maximizing $l_c = \sum_1^{r+1} l_k$ where l_k is defined in (7) .

Table 1 shows the bias and mean square error (MSE) for $\tilde{\alpha}$ and $\tilde{\beta}$ and also the bias and MSE for $\hat{\alpha}_c$ and $\hat{\beta}_c$. The conditional MLE clearly have smaller bias than $\tilde{\alpha}$ and $\tilde{\beta}$ and thus would probably be preferred in applications.

6. Final Remarks

We have presented procedures that may be useful for estimating an upper bound on the hitting rate of the faults remaining in the system. There does not seem to be any published data where software testing counters have been used, although Nayak (1988) credits this testing method to Huang (1977). A few studies have used replications of a debugging sequence and multiple versions of a program to automate the process of error detection (e.g., Nagel, Scholz, and Skrivan, 1984).

The model in (1) and, more generally, any model in which the ξ_i are defined by $\xi_i = \lambda_i - \lambda_{i+1}$, assumes that the effects of correcting faults are additive. If for some input condition, two or more faults give an error in the output, then the additivity assumption does not hold. Although disjoint fault sets are more common, cases of a nonempty intersection have been observed experimentally (Nagel, Scholz, and Skrivan, 1984). The additivity assumption can be preserved by (*a*) labeling such faults as a single fault and recording only the hitting frequency of the union of the fault sets or (*b*) partitioning the union of these sets into disjoint regions and recording the hitting frequencies as if each region corresponds to a different fault. The former is a simpler and more practical method since the hitting frequencies of the common fault regions may tend to be small.

In analyzing data obtained by recapture sampling, it is important to use estimators that are consistent (i.e., converge in probability to the true parameters). Although consistency is not guaranteed for estimators based on an unconditional likelihood, we have shown that certain moment estimators, obtained in the context of a conditional likelihood, are consistent and asymptotically normal; it is likely that the conditional maximum likelihood estimators are also consistent. Our asymptotic evaluation of the estimators assumes that the number of detected faults is large. It is, therefore, of interest to determine whether consistency of these or other estimators will hold when conditioning on other quantities, such as the total number ($m + r$ where $m = \sum_1^r m_i$) of error occurrences.

7. APPENDIX

Before we give the proof of Theorem 2, we require the following moments, which can be obtained by using (8) where V_{2k} and V_{3k} are linear function of the conditionally independent Poisson random variables $\{M_{ik}\}$.

$$E[(V_{2k} - \mu_{2k})^4] = 3a_0^2 E(Y_k^2) + a_0\mu_{1k} + 6a_0^3 E[Y_k(Y_k - \mu_{1k})^2] + a_0^4 E[(Y_k - \mu_{1k})^4]$$

$$E[(V_{3k} - \mu_{3k})^4] = 3a_2^2 E(Y_k^2) + a_4\mu_{1k} + 6a_1^3 E[Y_k(Y_k - \mu_{1k})^2] + a_1^4 E[(Y_k - \mu_{1k})^4]$$

$$E[(V_{2k} - \mu_{2k})^2(V_{3k} - \mu_{3k})^2] = \phi a_0\mu_{1k} + \phi a_0^3 E[(Y_k - \mu_{1k})^3] + \phi a_0^2 \mu_{1k}^2$$

$$+ \phi a_0^3 \mu_{1k} E[(Y_k - \mu_{1k})^2] + (a_1^2/a_0^2) E[(V_{2k} - \mu_{2k})^4]$$

$$a_p = \sum_{i=1}^{k-1} (i-1)^p \xi_i \quad p = 0,1,2,3,4 \quad \phi = a_0^{-1}a_2 - a_0^{-2}a_1^2$$

The limits, as $k \to \infty$, of these moments are finite because (i) the Y_k are bounded and converge in distribution to a uniform distribution on the interval $(0, s)$, and (ii) for each p=0,1,2,3,4, $a_p = \alpha(1 - e^{-\beta}) \sum_{i=1}^{k-1} (i-1)^p e^{-\beta(i-1)}$ converges to a finite positive limit as $k \to \infty$. Therefore, $E(U_k^2)$ converges to a finite limit where U_k is defined in (10). Thus as r tends to infinity, the limit in (9) is zero, which proves Theorem 1.

References

[1] Huang, J.C. (1977), Error detection through program testing. In Current Trends in Programming Methodology, 2, Ed. R.T. Yeh, pp. 16-43. Prentice Hall, Engelwood Cliffs, N.J.

[2] Jelinski, Z. and Moranda, P.M. (1972). Software reliability research. In *Statistical* Computer Performance Evaluation, Ed. W. Freiberger, pp. 465-84. Academic Press, New York.

[3] Nagel, P.M., Scholz, F.W., and Skrivan, J.A. (1984). Software reliability; additional investigations into modeling with replicated experiments. NASA Langley Research Center, Hampton, VA, NASA Contractor Report 172378, June.

[4] Nayak, T.K. (1988). Estimating population size by recapture sampling. *Biometrika*, 75, 113-120.

[5] Serfling, R.J. (1980).Approximation Theorems of Mathematical Statistics. John Wiley & Sons, New York.

r		$\hat{\alpha}_c$	$\hat{\beta}_c$	$\tilde{\alpha}$	$\tilde{\beta}$
15	Bias	0.000066	0.005990	-0.011208	0.055903
	MSE	0.000196	0.003188	0.000295	0.006629
20	Bias	-0.000001	-0.000001	-0.007733	0.041084
	MSE	0.000207	0.002241	0.000276	0.004115
25	Bias	0.000057	0.000002	-0.004581	0.032183
	MSE	0.000103	0.000680	0.000106	0.001701
30	Bias	0.000000	0.000000	-0.002472	0.022305
	MSE	0.000074	0.000789	0.000088	0.001288

Table 1. Bias and mean square error (MSE) of the conditional maximum likelihood estimators $\hat{\alpha}_c$, $\hat{\beta}_c$ and the moment estimators $\tilde{\alpha}$, $\tilde{\beta}$ based on 1,000 simulations with $\alpha = 0.10$ and $\beta = 0.35$.