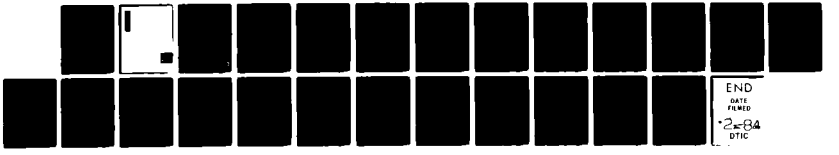
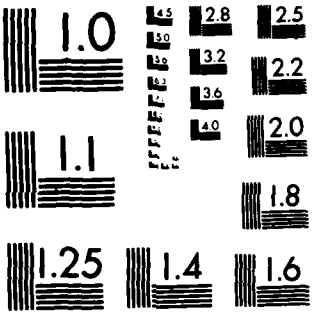
AD-A136 890 INSPECTIONS WITH UNKNOWN DETECTION PROBABILITIES(U) 1/1
STANFORD UNIV CA C DERMAN ET AL. NOV 83 TR-211
N00014-75-C-0561

UNCLASSIFIED F/G 12/1 NL





INSPECTIONS WITH UNKNOWN DETECTION PROBABILITIES

(12)

by

Cyrus Derman
Department of Civil Engineering
and Engineering Mechanics
Columbia University
New York, New York

and

Gerald J. Lieberman
Department of Operations Research
Stanford University
Stanford, California

and

Sheldon M. Ross
Department of Industrial Engineering
and Operations Research
University of California, Berkeley

TECHNICAL REPORT NO. 211

NOVEMBER 1983

SUPPORTED UNDER CONTRACT N00014-75-C-0561 (NR-047-200)
WITH THE OFFICE OF NAVAL RESEARCH

Gerald J. Lieberman, Project Director

DEPARTMENT OF OPERATIONS RESEARCH
AND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

INSPECTIONS WITH UNKNOWN DETECTION PROBABILITIES[1]
(The Proofreader Problem)

by

Cyrus Derman, Columbia University
Gerald J. Lieberman, Stanford University
Sheldon M. Ross, Univ. of Cal., Berkeley

1. Introduction and Summary

In this document we assume that

Suppose in an acceptance sampling situation the lot is subject to 100% inspection. The probability that a defective unit is detected is different for each inspector and is unknown. It is of interest to estimate N, the number of defective units in the lot (presumably, a decision to reject or accept the lot would be based on the estimate of N). Or, suppose satellites are used for surveillance over a given part of the earth with the detection of certain types of installations being the mission of a given satellite. However, for various reasons, it can be assumed the detection of any existing installation is uncertain with an unknown probability of detection that varies among satellites. The problem is to estimate the total number of installations based on the number observed. A third situation involves the reading of a manuscript by many proofreaders. Based on the results, it may be of interest to estimate the total number of typographical errors.

For purposes of exposition we shall, in formulating the model, use language suggested by the proofreader situation.

---

[1]This research has been partially supported by a) the U.S. Office of Naval Research under Contracts N00014-75-C-0561 with Stanford University and, b) the U.S. Air Force Office of Scientific Research (AFSC), USAF, under Grant AFOSR-81-0122 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

The proofreader problem has been treated: Polya [3] and Jewell [2]. In the context of wildlife recapture census, the literature reaches back to the 1950's (a reference list appears in G.A.F. Seber [4]).

We develop models for estimating the quantities of interest. Our models are generalizations of what has appeared in the wildlife recapture census and proofreading literature. In the context of the proofreading model the existing literature has considered the situation where all $K(K > 1)$ readers read the entire manuscript. In our models we allow for the possibility that the manuscript can be divided into several chapters. Each reader reads one or more, but not necessarily all, chapters. We look at this generalization in two ways. The first model is multi-variate with an unknown number of errors in each chapter to be estimated. The second model assumes that the number of errors in the entire manuscript has a Poisson distribution with unknown mean and that the relative sizes of the chapters are known. We rely on the method of maximum likelihood for estimating the unknown parameters. Typically the maximum likelihood estimates of the quantities are solutions to equations which must be solved numerically. In this paper we are not concerned with the statistical properties of these estimates. We are primarily concerned with the convergence properties and performance of an intuitive iterative procedure which, given the present generation of personal computers, can provide the desired numerical estimates in a matter of seconds.

2. GENERAL MODEL

We assume a "manuscript" with M, $M \geq 1$, "chapters" and K, $K > 1$ "proofreaders". Each proofreader is assigned a number of chapters to

read. Let K_i denote the set of proofreaders assigned to read chapter i, $i = 1, \ldots, M$; let L_j denote the set of chapters assigned to proofreader j, $j = 1, \ldots, K$; let N_i denote the unknown number of "errors" in chapter i; let p_j denote the unknown probability of proofreader j detecting a given error when he "reads" it. We assume independence from error to error so that the number of errors proofreader j, $j = 1, \ldots, K$ finds are independent binomial random variables with parameters $\sum_{i \in L_j} N_i$ and p_j, $j = 1, \ldots, K$.

Let

$$Q_i = \prod_{j \in K_i} (1-p_j)$$

be the probability that a given error in chapter i will not be found by any proofreader. Let $n(j,i)$ denote the number of errors that proofreader j finds in chapter i; let T_i denote the total number of different errors found in chapter i by all of the proofreaders assigned to read that chapter. The likelihood function of the observed data given $(N,p) = (N_1, \ldots, N_M, p_1, \ldots, p_K)$ is given by

$$L(\text{data} \mid (N,p)) = \prod_{i=1}^{M} \frac{N_i!}{(N_i-T_i)!d_i!} Q_i^{N_i-T_i} \prod_{j \in K_i} p_j^{n(j,i)} (1-p_j)^{T_i-n(j,i)}$$

(1)

$$= \prod_{i=1}^{M} \frac{N_i!}{(N_i-T_i)!d_i!} Q_i^{N_i} \prod_{j=1}^{K} \left(\frac{p_j}{1-p_j}\right)^{n(j)} ,$$

where d_i is a function of the data associated with chapter i and does not depend on (N,p) and where $n(j) = \sum_{i \in L_j} n(j,i)$ is the total number of errors found by proofreader j.

If we approximate by assuming the N_i's to be continuous variables and substitute $\log N$ for $d \frac{\log N!}{dN}$, on partial differentiation of $\log L$ with respect to the N_i's and p_j's we obtain equations that the maximum likelihood estimators $\hat{N}_i$, $\hat{p}_j$ of N_i and p_j must satisfy:

$$\hat{N}_i = \frac{T_i}{1-\hat{Q}_i}, \quad i = 1, \ldots, M \quad , \tag{2}$$

and

$$\hat{p}_j = \frac{n(j)}{\sum_{i \in L_j} \hat{N}_i} \quad , \quad j = 1, \ldots, K \quad , \tag{3}$$

where

$$\hat{Q}_i = \prod_{j \in K_i} (1-\hat{p}_j) \quad .$$

Combining (2) and (3) yields the equations

$$\hat{N}_i = \frac{T_i}{1 - \prod_{j \in K_i} \left(1 - \frac{n(j)}{\sum_{v \in L_j} \hat{N}_v}\right)}, \quad i = 1, \ldots, M \quad . \tag{4}$$

In addition to (4), we have the additional constraints that $\hat{N}_i \geq T_i$, $i = 1, \ldots, M$, which implies, also, that

$$\sum_{v \in L_j} \hat{N}_v \geq n(j), \quad j = 1, \ldots, K \ .$$

For given values $\{N_i(0), i = 1, \ldots, M\}$ define for $u \geq 1$,

$$N_i(u+1) = \frac{T_i}{1 - \prod_{j \in K_i} \left(1 - \frac{n(j)}{\sum_{v \in L_j} N_v(u)}\right)} \ , \quad i = 1, \ldots, M \ .$$

The above defined iteration is suggested by (2), where initially a value for $\hat{Q}_i$ is given which in turn generates a value for $\hat{N}_i$ which in turn generates another value for $\hat{Q}_i$, etc.

Proposition 1: If $N_i(0) = T_i$, $i = 1, \ldots, M$, then $\{N_i(u), u = 0, \ldots\}$ is non-decreasing in u for every i.

Proof: If $N_i(0) = T_i$, $i = 1, \ldots, M$, it is clear from (4) that $N_i(1) \geq N_i(0)$, $i = 1, \ldots, M$. However, replacing $N_i(0)$ by $N_i(1)$ increases the right side of (4) which means that $N_i(2) \geq N_i(1)$, $i = 1, \ldots, M$. Continuing, the proposition follows.

The monotonicity of Proposition 1 does not guarantee that $\lim_{u \to \infty} N_i(u)$ exists; i.e., we could have $N_i(u) \uparrow \infty$. Proposition 1 starts from the lowest possible value of N_i. The following proposition starts from a high value of N_i and asserts monotonicity in the opposite direction.

Proposition 2: Suppose $N_i(0) = T_i N$. If

$$\sum_{j \in K_i} \frac{n(j)}{\sum_{v \in L_j} T_v} > 1, \quad i = 1, \ldots, M \ , \tag{5}$$

then there exists N large enough such that $\{N_i(u), u = 0, 1, \ldots\}$ is non-increasing for every i. Consequently, $\{N_i(u), i = 1, \ldots, M; u = 0, \ldots\}$ converges to a solution of (4).

Proof: For N large enough, the left side of (5) is the dominating term, for each i, in the denominator of (4). For N large enough one gets that $N_i(1) \leq N_i(0)$, $i = 1, \ldots, M$. By the same argument used in the proof of Proposition 1, we get that $N_i(u+1) \leq N_i(u)$, $i = 1, \ldots, M$. Since the $N_i(u)$'s are bounded below by T_i for every i the sequences must each have a limit. That the limit satisfies (4) follows by the continuity of the functions involved in (4).

What still is an open question is whether, or under what conditions, (4) has a unique solution in the region $N_i \geq T_i$. When uniqueness can be established then that limit arrived at in Proposition 2 can be taken to be $\hat{N}_i$, $i = 1, \ldots, M$, and at the same time yielding the values $\hat{p}_j$, $j = 1, \ldots, K$.

Remark: There is a simple heuristic argument that also leads to the estimators provided by (4). As

$$N_i = T_i + \text{Number of errors missed in chapter } i$$

we obtain upon taking expectation that

$$N_i = E(T_i) + N_i \prod_{j \in K_i} (1-p_j).$$

Now given N_i, $i = 1, \ldots, M$, a natural estimate of p_j is the number of errors j finds divided by the number of errors in the chapters read by j, that is,

$$p_j \approx \frac{n(j)}{\sum_{v \in L_j} N_v} .$$

Hence, we see that

$$N_i \approx \frac{T_i}{1 - \prod_{j \in K_i} \left(1 - \frac{n(j)}{\sum_{v \in L_j} N_v}\right)} , \quad i = 1, \ldots, M .$$

3. Special Cases

(a) $M = 1$, $K > 1$. This is the case where all K proofreaders read the entire manuscript. This is the case that has been in the wild life recapture census literature (see Seber [4]) and more recently by Polya [3] for the case $K = 2$, Jewell [2] for $K > 2$. Equation (4), with $\hat{N} = N_1$, becomes the single classical equation

$$\hat{N} = \frac{T}{1 - \prod_{j=1}^{K} \left(1 - \frac{n(j)}{\hat{N}}\right)} \tag{6}$$

where $T = T_1$.

It is known (also see Corollary 1 to Proposition 4 below) that if $\max_j\{n(j)\} < T < \sum_{j=1}^{K} n(j)$, then (6) has a unique root in the interval $[T, \infty]$; if $\sum_{j=1}^{K} n(j) = T$ then $\hat{N} = \infty$ and if $T = \max_j\{n(j)\}$, then $\hat{N} = T$. If $\sum_{j=1}^{K} n(j) > T$ then condition (5) holds and both Propositions 1 and 2 apply. Let $\hat{N}$ be the unique finite root to (6). Since the right member of (6) is greater than the left member at $N = T$ (assuming $\max_j\{n(j)\} < T$) and is less when N is large enough (assuming

$\sum_{j=1}^{K} n(j) > T$) the curve defined by the right member crosses the line defined by the left exactly once at $N = \hat{N}$ from above. Thus, viewing the iterative procedure graphically, we see that $\{N(u)\} \uparrow$ whenever $N(0) < \hat{N}$ and $\{N(u)\} \downarrow$ whenever $N(0) > \hat{N}$. Since $\{N(u)\}$ would cease to be monotone if it crossed $\hat{N}$, the increasing sequence as well as the decreasing sequence must converge. The only point they can converge to is $N = \hat{N}$.

Suppose $\{n'(j)\}$ such that $\max\{n'(j)\} < T < \sum_{j=j}^{M} n'(j)$. Let $\hat{N}'$ be the root of (6) when $n'(j)$ replaces $n(j)$, $j = 1, \ldots, M$. We have

<u>Proposition 3</u>: If $\prod_{j=1}^{K} (1 - n'(j)/N) \geq \prod_{j=1}^{K} (1 - \frac{n(j)}{N})$ for every $N \geq T$, then $\hat{N}' \geq \hat{N}$.

<u>Proof</u>: Let $N(0) = \hat{N}'$. Then

$$\hat{N}' = \frac{T}{1 - \prod_{j=1}^{K} (1 - \frac{n'(j)}{\hat{N}'})}$$

$$\geq \frac{T}{1 - \prod_{j=1}^{K} (1 - \frac{n(j)}{\hat{N}'})}$$

$$= N(1) \quad .$$

That is, $N(1) \leq N(0)$, implying $N(u) \downarrow \hat{N}$; hence, $\hat{N} \leq \hat{N}'$.

If it is assumed that $p_1 = p_2 = \cdots p_K = p$, then the likelihood function becomes, since $M = 1$,

$$L(\text{data} \mid N,p) = \frac{N!}{(N-T)!d!} p^{\sum_{j=1}^{K} n(j)} (1-p)^{KN - \sum_{j=1}^{K} n(j)} \quad .$$

The equations for the maximum likelihood estimates are the same except that

$$\hat{p}_j = \hat{p} = \frac{\sum_{j=1}^{K} n(j)}{KN} \quad , \quad j = 1, \ldots, M \ .$$

Thus, under the assumption of equal p_i's equation (6) will be the same with $n'(j)$ replacing $n(j)$ where $n'(j) = \sum_{j=1}^{K} \frac{n(j)}{K}$. Now proposition 3 applies since $\prod_{j=1}^{K} (1 - \frac{n'(j)}{N}) \geq \prod_{j=1}^{K} (1 - \frac{n(j)}{N})$ for ever $N \geq t$; this follows from the convexity of $\log(1-x)$ in the interval $0 \leq$. Thus, for the same data, the assumption of equal p_j's always leads to a larger estimate of N.

Asymptotic variance and bias for the estimator $\hat{N}$ can be found in Darroch [1].

(b) One chapter is read by all proofreaders, all other chapters are read by only one proofreader. Here $K+1 = M > 1$; $i = 0, \ldots, M-1$; all proofreaders read chapter 0 and only proofreader j reads chapter j, $j = 1, \ldots, K$.

The equations (4) become

(7)
$$\hat{N}_0 = \frac{T_0}{1 - \prod_{j=1}^{K} \left(1 - \frac{n(j,0)+T_j}{\hat{N}_0 + \hat{N}_j}\right)}$$

$$\hat{N}_i = \left(\frac{\hat{N}_0 + \hat{N}_i}{n(i,0)+T_i} \right) T_i \quad , \quad i = 1, \ldots, M-1 \ .$$

The second part of (7) is equivalent to

$$\hat{N}_i = \frac{T_i}{n(i,0)} \hat{N}_0 \quad , \quad i = 1, \ldots, M-1 \quad . \tag{8}$$

Substituting (8) in the first part of (7) yields

$$\hat{N}_0 = \frac{T_0}{1 - \prod_{j=1}^{K} \left(1 - \frac{n(j,0)+T_j}{\hat{N}_0(1+T_j/n(j,0))}\right)} \tag{9}$$

$$= \frac{T_0}{1 - \prod_{j=1}^{K} \left(1 - \frac{n(j,0)}{\hat{N}_0}\right)} \quad , \tag{9}$$

the classical equation discussed in special case (a) with $T = T_0$ and $n(j) = n(j,0)$. Thus, (9) has a unique solution $\hat{N}_0$ which can be obtained by iteration, and once $\hat{N}_0$ is obtained, $\hat{N}_i$, for $i = 1, \ldots, K$, follows by (8).

4. Results of Simulations

In general when $M > 1$ the usefulness as an estimate of the N_i's of whatever limits result from use of the iterative procedure is in question since uniqueness in the region $\hat{N}_i \geq T_i$, $i = 1, \ldots, M$ has not been demonstrated. Neither has any results pertaining to the speed of convergence been shown. To see what is likely to be the case some experiments were simulated for several cases. In each case convergence to a unique and likely value of N_i appears to occur and the convergence

takes place in a matter of seconds when using a modern personal computer. In each case we initiated two sets of calculations - one starting with $N_i(0) = T_i$, the other with $N_i(0) = (T_i + C)$ where C was large enough to produce a decreasing sequence. In each case the calculations lead rapidly to the same values for $\hat{N}_i$, $i = 1, \ldots, M$. Specifically, we let $M = K = 4$. We had two different chapter-assignment designs:

$$D_1 = \begin{Bmatrix} 1110 \\ 1101 \\ 1011 \\ 0111 \end{Bmatrix} \quad , \quad D_2 = \begin{Bmatrix} 1100 \\ 0110 \\ 0011 \\ 1001 \end{Bmatrix}$$

where a 1 or 0 occurs in entry a_{ji} of the design matrix according to whether or not reader j is assigned to read chapter i. We also had 5 different probability $\{p_j\}$ combinations for each design:

	Readers			
Combinations	1	2	3	4
1	.90	.90	.90	.90
2	.10	.15	.75	.80
3	.10	.15	.20	.25
4	.10	.15	.20	.75
5	.60	.70	.70	.80

In every case we set $N_i = 70$, $i = 1, \ldots, 4$ and $C = 100$. We take as the estimates of N_i the nearest integer to the limits of the iterative procedure. The number of iterations required to reach the estimate was taken to be the number of iterations until the nearest integer was

reached. In practice, more iterations are used in order to recognize when the procedure appears to converge. However, the length of real time required turns out to be negligible. The results of the experiments are summarized in Table 1.

As would be expected the accuracy of the estimates improves with increasing p_i's. This would be expected intuitively and from the formula for the asymptotic variance of $\hat{N}$ given by Darroch [1] for the case of $M = 1$. The number of iterations required also appears to decrease with increasing p_i.

5. Poisson Model

Assume that the ratio of the size of chapter i to the whole manuscript is α_i, $\alpha_i \geq 0$, $\sum_{i=1}^{M} \alpha_i = 1$. Assume the number of errors $(N_1, \ldots, N_M)$ are independent random variables with a Poisson distribution having mean $\alpha_i\lambda$, $i = 1, \ldots, M$ where λ, as opposed to the α_i, is unknown. Under this assumption, following Jewell [2], the likelihood function, averaging (1) over the possible value of $\{N_1, \ldots, N_M\}$, becomes

$$(10) \qquad L(\text{data} \mid (\lambda,p)) = D \prod_{i=1}^{M} (Q_i \times \lambda)^{T_i} e^{-\lambda \sum_{i=1}^{M} \alpha_i (1-Q_i)} \prod_{j=1}^{K} \left(\frac{p_j}{1-p_j}\right)^{n(j)}$$

where D is a function of the data. Taking partial derivatives of log L, we see that the maximum likelihood estimate $\hat{\lambda}$, $\hat{p}_i$ of λ and p_i must satisfy

Table 1

	Combinations	Chapters 1 D_1	1 D_2	2 D_1	2 D_2	3 D_1	3 D_2	4 D_1	4 D_2	Estimates of p_j's Readers 1 D_1	1 D_2	2 D_1	2 D_2	3 D_1	3 D_2	4 D_1	4 D_2
1	$\hat{N}_i$	70	71	70	70	70	71	70	71	.95	.92	.90	.92	.90	.88	.90	.91
	No. of Iters. from T_i	1	2	1	1	1	1	1	1								
	No. of Iters. from T_i + 100	3	4	3	6	3	5	3	5								
2		67	74	73	64	71	78	72	72	.09	.09	.13	.13	.73	.65	.81	.81
		4	12	4	10	3	10	3	9								
		10	15	9	19	10	17	9	14								
3		116	33	86	34	88	66	87	46	.07	.16	.10	.19	.12	.23	.20	.38
		31	20	38	22	35	25	35	26								
		>40	>40	32	>40	34	>40	33	37								
4		80	64	64	80	73	83	62	70	.09	.11	.13	.13	.19	.17	.76	.79
		19	25	11	39	12	26	21	21								
		19	37	26	37	22	>40	18	>40								
5		71	69	71	75	69	74	71	72	.64	.51	.70	.65	.73	.77	.76	.78
		1	3	2	7	2	4	1	5								
		5	12	4	8	4	8	5	7								

$$(11) \qquad \hat{\lambda} = \frac{\sum_{i=1}^{M} T_i}{1 - \sum_{i=1}^{M} \alpha_i \hat{Q}_i}$$

$$\hat{p}_j = \frac{n(j)}{\hat{\lambda} - \sum_{v \notin L_j} T_v} \quad , \quad j = 1, \ldots, M \quad ,$$

leading to the equation for $\hat{\lambda}$ being a solution of the equation

$$(12) \qquad \hat{\lambda} = \frac{\sum_{i=1}^{M} T_i}{1 - \sum_{i=1}^{M} \alpha_i \prod_{j \in K_i} \left(1 - \frac{n(j)}{\hat{\lambda} - \sum_{v \notin L_j} T_v}\right)} \quad .$$

Thus, where the model in Section 2 leads to M equations in M variables, the Poisson model reduces the number of equations to one equation with one variable. For case (a) in Section 3 (12) reduces to (6) as was pointed out by Jewell [2]. The same numerical iteration suggests itself. If $\lambda(0)$, the initial value of the iteration is $\sum_{i=1}^{M} T_i$, one has the same monotonicity of Proposition 1. Numerical calculations indicate that the sequence $\{\lambda(u)\}$ will converge. The question of uniqueness of the solution to (12) in the region $\lambda \geq \sum_{i=1}^{M} T_i$ is open. (Since $n(j) \leq \sum_{v \in L_j} T_v$, $\lambda \geq \sum_{i=1}^{M} T_i$ implies $n(j) \leq \lambda - \sum_{v \notin L_j} T_v$.) The next provides a partial answer to this question.

Let

$$c_j = \sum_{v \notin L_j} T_v \quad , \quad j = 1, \ldots, K$$

and

$$T = \sum_{v=1}^{M} T_v \quad .$$

Proposition 4: If for every i, $i = 1, \ldots, M$

$$\sum_{\substack{k \neq j \\ k \in K_i}} n(k) \geq \frac{2c_j}{1-c_j/T} \quad , \quad \forall\, j \in K_i \quad , \tag{13}$$

then there is a most one solution to (12) in $[T,\infty)$.

Proof: Invert both sides of (12) letting $z = 1/\lambda$ to get

$$z = G(z)/T \tag{14}$$

where

$$G(z) = 1 - \sum_{i=1}^{M} \alpha_i \prod_{j \in K_i} \left(1 - \frac{n(j)z}{1-c_j z}\right) \quad .$$

We shall show that (13) is a sufficient condition for $G(z)$ to be a concave function in the interval $[0,1/T]$. To this end, the first derivative with respect to z is

$$G'(z) = - \sum_{i=1}^{M} \alpha_i H_i'(z)$$

where

$$H_i(z) = \prod_{j \in K_i} \left(1 - \frac{n(j)z}{1-c_j z}\right) \quad , \quad i = 1, \ldots, M \quad .$$

But

$$H_i'(z) = - \sum_{j \in K_i} P_j(z)$$

where

$$P_j(z) = \prod_{k \neq j} \left(1 - \frac{n(k)z}{1-c_k z}\right) \frac{n(j)}{(1-c_j z)^2} \quad , \quad j \in K_i \quad .$$

Thus,

$$H_i''(z) = - \sum_{j \in K_i} P_j'(z) \quad ,$$

but

$$P_j'(z) = \frac{2n(j)}{(1-c_j z)^3} \prod_{k \neq j} \left(1 - \frac{n(k)z}{1-c_k z}\right) - \frac{n(j)}{(1-c_j z)^2} \sum_{k \neq j} \prod_{v \neq k,j} \left(1 - \frac{n(v)z}{1-c_v z}\right) \frac{n(k)}{1-c_k z}$$

$$= \frac{n(j)}{(1-c_j z)^2} \prod_{v \neq j} \left(1 - \frac{n(v)z}{1-c_v z}\right) \left\{ \frac{2c_j}{1-c_j z} - \sum_{k \neq j} \frac{n(k)}{(1-(c_k + n(k))z)(1-c_k z)} \right\} \quad .$$

Thus, the sign of $P_j'(z)$ is the sign of the expression in brackets which is less than or equal to $\frac{2c_j}{1-c_j/T} - \sum_{k \neq j} n(k)$. Then, by the assumption (13), the above is non-positive in $[0,1/T]$ for every $j \in K_i$. Therefore, $H_i''(z) \geq 0$ in $[0,1/r]$ for every i, and consequently,

$$G''(z) \leq 0 \quad , \quad 0 \leq z \leq \frac{1}{T} \quad ,$$

as was to be shown.

In the case where every proofreader reads every chapter, $c_j = 0$, $j = 1, \ldots, K$ (or we can consider $M = 1$, and (12) becomes (6)), and we have the already known result.

Corollary 1: If every proofreader reads every chapter, then (12) has at most one solution in the interval $\hat{\lambda} \geq T$.

Proof: Condition (13) is satisfied in this case.

When $G(z)$ is concave, equation (12) will have exactly one solution in the interval $\hat{\lambda} \geq T$ if $G'(0)/T > 1$. Since

$$G'(0) = -\sum_{i=1}^{M} \alpha_i H_i'(0)$$

$$= \sum_{i=1}^{M} \alpha_i \sum_{j \in K_i} P_j(0)$$

$$= \sum_{i=1}^{M} \alpha_i \sum_{j \in K_i} n(j) \quad ,$$

we have

Corollary 2: If (13) holds and

$$\sum_{i=1}^{M} \alpha_i \sum_{j \in K_i} n(j) > T \quad ,$$

then there exists exactly one solution to (12) in the interval $\hat{\lambda} \geq T$.

Even if (12) does not have a unique solution we can bound all the roots of (12). Let $\tilde{\lambda}$ be the unique solution to (12) in $[T,\infty)$ when $c_j = 0$, $j = 1, \ldots, K$. We have

Corollary 3: Let $\hat{\lambda}$ be any solution to (12) in $[T,\infty)$, then $\hat{\lambda} \leq \tilde{\lambda}$.

Proof: This corollary follows from the fact that the right side of (12) is continuous and that it is always largest when $c_j = 0$, $j = 1, \ldots, K$.

The previous proposition and corollaries deal with sufficient conditions for a unique solution to (12) to exist and an upper bound $\tilde{\lambda}$ on the possible solutions in case there may be more than one solution. The following proposition indicates that the iterative method for solving (12) indicates when a unique solution exists or at worst provides bounds within which the appropriate solution to (12) exists. Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_r$ be the roots of (12) in $[T,\infty)$.

Proposition 5: If $\lambda(0) = T$, then $\lim_{u\to\infty} \lambda(u) = \lambda_1$; if $\lambda(0) = \tilde{\lambda}$ then $\lim_{u\to\infty} \lambda(u) = \lambda_r$. Consequently, if the two limits are equal, then (12) has a unique solution in $[T,\infty)$; otherwise $T \leq \lambda_1 \leq \hat{\lambda} \leq \lambda_r \leq \tilde{\lambda}$.

Proof: Let $\phi(\lambda)$ denote the right number of (12). We shall exploit the fact that $\phi(\lambda)$ is increasing in λ and that if $\lambda(0) = T$, $\{\lambda(u)\}$ is increasing in u. Suppose $\lim_{u\to\infty} \lambda(u) = \bar{\lambda} > \lambda_1$. Then there exists a u such that $\lambda(u) < \lambda_1$ and $\lambda(u+1) > \lambda_1$. That is, we have

$$\begin{aligned} \phi(\lambda(u)) &= \lambda(u+1) \\ &\geq \lambda_1 \\ &= \phi(\lambda_1) \quad , \end{aligned}$$

a contradiction of the fact that $\phi(\lambda(u)) < \phi(\lambda_1)$. The proof is analogous for $\lambda(0) = \tilde{\lambda}$.

5. Remarks:

In the cases where we have not shown convergence of the iterative procedure, numerical examples have shown convergence quite often and in a matter of seconds on a personal computer. However, if a personal computer is not available or convergence may not occur, a two-step iterative procedure suggests itself. In equations (2) the estimate of N_i would be immediate if we had an estimate of $\hat{Q}_i$. If we go beyond the sufficient statistics $\{T_i, n(j)\}$ such estimates are available. Let S_i denote the number of distinct errors found by all proofreaders other than proofreader i in the chapters read by proofreader i. Within this set of distinct errors, let s_i denote the number of errors found by proofreader i. Then an estimate $\tilde{p}_i$ of p_i is given by s_i/S_i. Inserting $\tilde{p}_i$ for the p_i's in Q_i yields an estimate of Q_i from which an estimate $\tilde{N}_i$ of N_i follows from (2). The procedure could terminate at this point or, perhaps, be carried on one more iteration by setting $N_i(0) = \tilde{N}_i$ in (4) and then letting $N(1)$ be the final estimate N_i.

A model not covered in the paper would be of interest. Errors may fall into different categories where the p_j's for each reader would vary in an unknown way with each category. If the categories are recognizable, then the present model can be adapted - treating each category separately. However, if the categories are not recognizable this device will not work.

Acknowledgement:

We are indebted to Michael I. Lieberman for his computer programming efforts required to produce the simulation results of Section 4.

References

[1] Darroch, J. N., The Multiple-Recapture Census I. Estimation of a Closed Population, Biometrika, 1958, 343-359.

[2] Jewell, W. S., Estimating Undetected Errors, Technical Report, Univ. of California, Berkeley, February 1983.

[3] Polya, G., Probabilities in Proofreading, American Math. Monthly 83, 42.

[4] Seber, G.A.F., The Estimation of Animal Abundance, 1973, Hafner Press, N.Y.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 211	2. GOVT ACCESSION NO.: AD-A136 890	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): INSPECTIONS WITH UNKNOWN DETECTION PROBABILITIES		5. TYPE OF REPORT & PERIOD COVERED: Technical Report
		6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): C. DERMAN, G. J. LIEBERMAN, AND S. M. ROSS		8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0561
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Operations Research and Department of Statistics - Stanford University, Stanford, California 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-047-200)
11. CONTROLLING OFFICE NAME AND ADDRESS: Operations Research, Code 434, Office of Naval Research, Arlington, Virginia 22217		12. REPORT DATE: November 1983
		13. NUMBER OF PAGES: 20
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report): UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

acceptance sampling
inspection
satellite surveillance
maximum likelihood
proofreading errors
Poisson

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(see next page)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT: Inspections with Unknown Detection Probabilities

by C. Derman, G. J. Lieberman, and S. M. Ross

Suppose in an acceptance sampling situation the lot is subject to 100% inspection. However, the inspector is not perfect so that the probability of a defective unit being detected is unknown. It is of interest to estimate N, the number of defective units in the lot (presumably, a decision to accept or reject the lot is based on the estimate of N). Or suppose satellites are used for surveillance over a given part of the earth. The detection of certain types of installations is the mission of a given satellite. However, for various reasons, it can be assumed the detection of any existing installation is uncertain with unknown probability of detection. It may be of interest to estimate the total number of installations based on the number observed. A third situation involves the proofreading of a manuscript. Based on the proofreading of the manuscript, it may be of interest to estimate the total number of typographical errors. There is a long literature, dating back to 1500, on special cases of these problems.

General models are developed for estimating the quantities of interest. An iterative scheme for calculating the maximum likelihood estimator is proposed, and convergence properties are proven for special cases.


DATE
ILME