MEMORANDUM
RM-5360-ISA/ARPA
JULY 1967

ESTIMATING FROM MISCLASSIFIED DATA

S. J. Press

PREPARED FOR:
THE OFFICE OF THE ASSISTANT SECRETARY
OF DEFENSE/INTERNATIONAL SECURITY AFFAIRS
AND THE
ADVANCED RESEARCH PROJECTS AGENCY

The RAND Corporation
SANTA MONICA • CALIFORNIA

MEMORANDUM
RM-5360-ISA/ARPA
JULY 1967

ESTIMATING FROM MISCLASSIFIED DATA

S. J. Press

This research is supported by the Department of Defense under Contract DAHC15 67 C 0158, monitored by the Assistant Secretary of Defense (International Security Affairs), and Contract DAHC15 67 C 0143, monitored by the Advanced Research Projects Agency. Any views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the Office of the Assistant Secretary of Defense (International Security Affairs) ~~or of~~ the Advanced Research Projects Agency.

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA • 90406

Published by The RAND Corporation

PREFACE

The research summarized in this Memorandum provides a method for estimating the proportions of items in each of several categories, based upon an item-by-item classification in which many items may be misclassified.

The motivation for the work had its genesis in a desire to compensate for incorrect answers that might be found in prisoner-of-war interviews. In that context, the items being classified are subjects in an interview (interrogation), and the misclassification takes place when the subject either deliberately lies, or for some other reason, his answer does not correspond to the facts.

The technique developed is statistical, and may be applied to a wide variety of problems, both military and nonmilitary. In such problems, it is desired to determine the characteristics of a group of people or items in which large scale misclassification is inherently a factor. The research should be of interest to planners and evaluators of military surveys, directors of counterinsurgency programs, and operations

The author, a consultant to the RAND Corporation, is Associate Professor at the Graduate School of Business, University of Chicago.

SUMMARY

The problem examined in this Memorandum is that of estimating the category probabilities for items which may have been misclassified. A specific case of interest is that in which the items are subjects being interviewed and the subjects may be hostile. A hostile subject is one whose response to a question posed in an interview does not correspond to the true or factual situation. Although some surveys contain very few hostile subjects, others contain hostility as an inherent factor, such as those in which the members of a surveyed group have reason to deliberately mislead the interviewer. Of course the misclassified items could equally well be diagnosed patients in a hospital, accused parties in a court, or any one of many possible constructs. The theory is developed in a subject-interview context, however, in order to be specific.

The procedure recommended for this problem requires that an assessment be made, for each subject, of the probability that that subject is hostile. These probabilities are then combined with the actual responses to yield maximum likelihood estimators of the parameters. The problem is reduced to one of concave programming with a logarithmic objective function. Efficiency of the estimators is discussed in terms of their variances. A Bayesian approach to evaluation of the misclassification probabilities (or hostility probabilities) is presented and an opposition strategy is offered. Numerical examples are given to illustrate application of the procedure and the effect of ignoring the misclassified items problem.

ACKNOWLEDGMENTS

The author has benefited from conversations with various people at The RAND Corporation, the University of California, Los Angeles, and the University of Chicago. However, L. H. Wegner of RAND and William Kruskal of the University of Chicago were especially helpful.

CONTENTS

I. INTRODUCTION

In sampling from a human population it is usually assumed that the interviewee is cooperative and that his responses to questions correspond to the true situation with respect to this individual. Thus, in a sample of size n with S_n individuals who claim to have a given characteristic, S_n/n is the maximum likelihood estimator of the population proportion corresponding to the given characteristic. If, however, some of the people questioned are hostile to the interviewer in that their responses are false for some reason, S_n/n is no longer a reasonable estimator, and a new estimation procedure must be found.

In this Memorandum procedures are developed for estimating the true population proportions in each category vis-à-vis a noncooperative group of interviewees. The two-category and the multicategory response cases are treated separately, owing to the intrinsic interest of the two-category response case. Maximum likelihood estimators are developed for both cases, and it will be seen that to evaluate the estimator explicitly for a sample of n subjects and r categories requires solution of a simple concave programming problem involving a logarithmic objective function in variables confined to the unit interval. Standard gradient methods for solving concave programming problems like this one are already available. For a survey containing many questions, the estimators would be evaluated for each question separately, and in the following it will be assumed that the analysis applies to just a single question. An outline will be given for generalizing the analysis to consideration of many questions simultaneously.

An approximate method for estimating the category proportions, which yields results rapidly for two categories, is outlined in Sec. IV.

The analysis to be developed is couched in the subject-interview framework only to be specific. In fact, the reader may equally well interpret the problem as one involving n items to be classified into one of r mutually exclusive categories, and although the classifier is not certain as to how to classify perfectly, he can assign probabilities of correctly classifying each of the items.

Previous investigations of this type of problem have been mostly concerned with methods of maximizing the number of cooperative interviewees. For example, a recent recommendation for increasing cooperation made by Warner[1] involves inducing the subject to be truthful by convincing him he is responding only according to a probabilistic mechanism.

In some situations there might be available a priori information on the probability distribution of the population proportion parameter for each of the categories. In such cases the Bayesian approach suggested by Hendricks[2] might reduce the error bias caused by the unreliable data.

In a different line of approach, Mote and Anderson[3] studied the effect of errors in classifying the subject on the usual chi-square tests of hypotheses. They found that, if the errors are ignored, the test size will increase and the asymptotic power will be reduced. Except for some special cases, however, the problem cannot be solved without knowledge of the misclassification probabilities.

The situation considered here assumes that it is possible for an assessment to be made of the "reliability" of the interviewee's responses (a Bayesian interpretation is given in Sec. IV). That is, it is assumed that it is possible to establish the probability that the subject's answer is correct. This information is then incorporated into the estimation procedure. It should be noted that the analysis to follow does not depend upon the reason the subject's response does not correspond to the true situation. Some subjects, of course, may deliberately lie. However, the responses of others may not correspond to the facts because of indigenous cultural differences between the interviewer and interviewee, psychological problems of the interviewee, semantic difficulties, etc. In the future, the terms "truth telling" and "lying" will be used to denote the extremes of discrepancy between the subject's response and the true situation. However, the terms should be interpreted in the general sense described above. (In case the context were not that of subjects being interviewed, the "items" might be misclassified for a wide variety of reasons, depending on the specific case at hand.)

Estimates obtained from sample surveys usually contain some bias that can be associated directly with data obtained from subjects who "lie," i.e., hostile subjects. The effect is of course small if the degree of hostility in the survey is proportionately small. However, in certain marketing, advertising, and voter preference surveys, hostility to the interviewer is inherent, and is therefore too large a factor to be ignored.

Section II contains the development of the maximum likelihood equations for various special cases of interest. Section III provides a discussion of the efficiency of the estimators for this problem and illustrates the estimation procedures with numerical examples. Section IV contains an analysis of the considerations that surround the problem of assessing the misclassification probabilities (and of ignoring them), and the effect of assessment errors on the results. Section V considers the problem of generalizing the analysis to include simultaneous evaluation of many questions. Finally, Sec. VI examines the problem from an entropy-information standpoint, and suggests an optimal strategy for hostile groups.

II. MAXIMUM LIKELIHOOD ESTIMATION

In this section the notation, terminology, and general model adopted throughout are introduced and the maximum likelihood estimation equations are developed. It is generally assumed that there are n subjects interviewed, any or all of whom may be hostile.

MULTICATEGORY RESPONSE CASE

Each question is phrased so that the response of every subject can be placed into one of r mutually exclusive and exhaustive categories. The interviewee is assigned a probability that his response is truthful (this assignment is based upon collateral information using procedures discussed in Sec. VI). Let

$$X_{kj} = \begin{cases} 1, & \text{if the jth subject } \underline{\text{actually}} \text{ belongs to category k} \\ 0, & \text{otherwise} \end{cases}$$

$$Z_{kj} = \begin{cases} 1, & \text{if the jth subject } \underline{\text{claims}} \text{ to belong to category k} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_j = \begin{cases} 1, & \text{if the jth subject tells the truth} \\ 0, & \text{otherwise} \end{cases}$$

for $k = 1, 2, \ldots, r$, $r \geq 2$; and $j = 1, 2, \ldots, n$.

Define the r-dimensional unit vectors $\mu_1' = (1, 0, \ldots, 0)$, $\ldots$, $\mu_r' = (0, \ldots, 0, 1)$. Let the result of the jth interview be denoted by the unit vector $Z_j' \equiv (Z_{1j}, \ldots, Z_{rj})$, and denote the true characterization by the vector $X_j' \equiv (X_{1j}, \ldots, X_{rj})$. Also, we will use lower case x's and z's to denote observed values. The assumptions

of the model may be summarized as

$$\text{(a)}\quad P\{X_j = \mu_k\} = p_k, \qquad k = 1, \ldots, r \tag{1a}$$

where the p_k are unknown parameters satisfying $0 \le p_k \le 1$, $\Sigma_{k=1}^{r} p_k = 1$, and for convenience, $q_k \equiv 1 - p_k$;

$$\text{(b)}\quad P\{Y_j = 1 \,|\, X_j = \mu_k\} \equiv \pi_{kj}^{(1)} \tag{1b}$$

where the $\pi_{kj}^{(1)}$ are all known, and of course $P\{Y_j = 0 \,|\, X_j = \mu_k\} = 1 - \pi_{kj}^{(1)}$;

$$\text{(c)}\quad P\{Z_j = \mu_k \,|\, Y_j = 1, X_j = \mu_k\} = 1 \text{ and}$$

$$P\{Z_j = \mu_k \,|\, Y_j = 0, X_j = \mu_m\} = \begin{cases} 0, & \text{if } k = m \\ \dfrac{1}{r-1}, & \text{if } k \neq m \end{cases} \tag{1c}$$

$$\text{(d)}\quad (Z_1, Z_2, \ldots, Z_n) \text{ are mutually independent random vectors} \tag{1d}$$

Note that although Assumption (c) implies that all categories are equally likely to be selected by a lying subject, the extension to a nonuniformly weighted distribution is immediate. That is, we could replace Assumption (d) by $P\{Z_j = \mu_k \,|\, Y_j = 0, X_j = \mu_m\} = \gamma_m$, if $k \neq m$, where the γ_m's are known constants satisfying $0 \le \gamma_m \le 1$, $\Sigma_1^{r-1} \gamma_m = 1$. For simplicity, assume $\gamma_m = (r-1)^{-1}$.

Now define

$$\pi_{kj}^{(0)} \equiv P\{Y_j = 1 \,|\, X_j \neq \mu_k\} \tag{2a}$$

It is seen that since

$$\pi_{kj}^{(0)} = \frac{P\{Y_j = 1, X_j \neq \mu_k\}}{P\{X_j \neq \mu_k\}}$$

$$\pi_{kj}^{(0)} = \frac{\sum_{\ell \neq k}^{r} \pi_{\ell j}^{(1)} p_\ell}{q_k} \tag{2b}$$

for $j = 1, \ldots, n$, $k = 1, \ldots, n$. Note from Eq. 2b that when $r = 2$, knowledge of $\pi_{1j}^{(1)}$, $\pi_{2j}^{(1)}$ implies knowledge of $\pi_{1j}^{(0)}$, $\pi_{2j}^{(0)}$. However, this is not true for $r > 2$.

The probability mass function for Z_j is similar to that of a multinomial, and is given by

$$P\{Z_j = z_j\} = \prod_{m=1}^{r} \left[P\{Z_j = \mu_m\}\right]^{z_{mj}} \tag{3}$$

with $z_j' \equiv (z_{1j}, \ldots, z_{rj})$, for $j = 1, \ldots, n$. To evaluate Eq. 3 it is necessary to determine the unconditional probability distribution of Z_j. Since

$$P\{Z_j = \mu_k\} = P\{Z_{kj} = 1\}$$

it is only necessary to consider the latter.

$$P(Z_{kj} = 1) = p_k P(Z_{kj} = 1 | X_{kj} = 1) + q_k P(Z_{kj} = 1 | X_{kj} = 0)$$

But

$$P(Z_{kj} = 1 | X_{kj} = 0) = P(Z_{kj} = 1 | X_{kj} = 0, Y_j = 1)\, P(Y_j = 1 | X_{kj} = 0)$$
$$+ P(Z_{kj} = 1 | X_{kj} = 0, Y_j = 0)\, P(Y_j = 0 | X_{kj} = 0)$$

or,

$$P(Z_{kj} = 1 | X_{kj} = 0) = \frac{1 - \pi_{kj}^{(0)}}{r - 1}$$

Similarly

$$P(Z_{kj} = 1 \mid X_{kj} = 1) = \pi_{kj}^{(1)}$$

Hence

$$P(Z_{kj} = 1) = p_k \pi_{kj}^{(1)} + \frac{q_k(1 - \pi_{kj}^{(0)})}{r - 1} \tag{4}$$

Substituting Eq. 4 into Eq. 3 gives

$$P\{Z_j = z_j\} = \prod_{m=1}^{r} \left[p_m \pi_{mj}^{(1)} + \frac{q_m(1 - \pi_{mj}^{(0)})}{r - 1} \right]^{z_{mj}}$$

for $j = 1, 2, \ldots, n$. Thus, if L^* denotes the logarithm of the likelihood function $L(z_1, \ldots, z_n \mid p_1, \ldots, p_r)$,

$$L^* = \ell n\, L = \sum_{j=1}^{n} \ell n\, P\{Z_j = z_j\}$$

or

$$L^* = \sum_{j=1}^{n} \sum_{m=1}^{r} z_{mj}\, \ell n \left[p_m \pi_{mj}^{(1)} + \frac{q_m(1 - \pi_{mj}^{(0)})}{r - 1} \right]$$

Now transform the parameters by letting $\theta_m = p_m$, $m = 1, 2, \ldots, r - 1$. Since $\Sigma_1^r p_m = 1$, $p_r = 1 - \Sigma_1^{r-1} \theta_m$. Substitution gives

$$L^* = \sum_{j=1}^{n} \left\{ \sum_{m=1}^{r-1} z_{mj}\, \ell n \left[\theta_m \pi_{mj}^{(1)} + \frac{(1 - \theta_m)(1 - \pi_{mj}^{(0)})}{r - 1} \right] \right.$$

$$\left. + z_{rj}\, \ell n \left[\left(1 - \sum_{m=1}^{r-1} \theta_m\right) \pi_{rj}^{(1)} + \sum_{m=1}^{r-1} \theta_m \frac{(1 - \pi_{rj}^{(0)})}{r - 1} \right] \right\} \tag{5a}$$

Next note from Eq. 4 that since the response of each subject must fall into one of the r categories,

$$\sum_{k=1}^{r} P\{Z_{kj} = 1\} = \sum_{k=1}^{r} \left[p_k \pi_{kj}^{(1)} + \frac{q_k (1 - \pi_{kj}^{(0)})}{r - 1} \right] = 1$$

However, it may be checked that since this result is implied by Eq. 2b, it is not really an additional constraint on the likelihood function. Substituting the constraint of Eq. 2b into the equation for L^* yields

$$L^* = \sum_{j=1}^{n} \sum_{m=1}^{r-1} z_{mj} \, \ell n \left\{ \theta_m \pi_{mj}^{(1)} + \frac{1 - \theta_m}{r - 1} \right.$$

$$\left. - \frac{1}{r - 1} \left[\sum_{\substack{\ell=1 \\ \ell \neq m}}^{r-1} \pi_{\ell j}^{(1)} \theta_\ell + \pi_{rj}^{(1)} \left(1 - \sum_{\ell=1}^{r-1} \theta_\ell \right) \right] \right\}$$

$$+ \sum_{j=1}^{n} z_{rj} \, \ell n \left\{ \left(1 - \sum_{m=1}^{r-1} \theta_m \right) \pi_{rj}^{(1)} + \sum_{m=1}^{r-1} \frac{\theta_m}{r - 1} \left(1 - \pi_{mj}^{(1)} \right) \right\} \qquad (5b)$$

The problem that must be solved is that of finding, for L^* defined in Eq. 5b,

$$\max_{\theta} L^*(\theta_1, \ldots, \theta_{r-1})$$

subject to the linear constraints that

$$0 \leq \theta_1, \theta_2, \ldots, \theta_{r-1} \leq 1 \qquad 0 \leq \sum_{1}^{r-1} \theta_i \leq 1$$

Recall that $\ell n \, x$ is a concave function of x, for any scalar x. Hence, it follows by definition that $\ell n \, (g'\theta + v)$ is concave in

$\theta \equiv (\theta_1, \ldots, \theta_{r-1})$ for any (g, v). Noting that L^* is just a sum of such functions, L^* is concave in θ. Thus, ours is a concave programming problem that may be solved (if the dimension of the problem is small enough) using any one of several standard computer routines. (See, for example, Rosen[4] for an exposition of the Gradient Projection method.) In any case, since the derivatives of L^* are fundamental to the solution, an idea of the magnitude of difficulty involved can be obtained by examining the unconstrained solution. The classical differentiation approach (neglecting the constraints) yields estimators $\hat{\theta}_k$ which satisfy the equations

$$\sum_{j=1}^{n} \sum_{m=1}^{r-1} \frac{z_{mj} v_{m\ell}}{v_m'\theta + (1 - \pi_{rj}^{(1)})/(r-1)} + \sum_{j=1}^{m} \frac{z_{rj} w_\ell}{w'\theta + \pi_{rj}^{(1)}}$$

$$\ell = 1, \ldots, r-1 \qquad (6)$$

where $v_{m\ell} = (\pi_{rj}^{(1)} - \pi_{\ell j}^{(1)})/(r-1) \qquad \ell \neq m$

$$v_{mm} = \pi_{mj}^{(1)} + (\pi_{rj}^{(1)} - 1)/(r-1)$$

$$v_m' = (v_{m1}, \cdots, v_{m,\, r-1})$$

$$w_\ell = -\pi_{rj}^{(1)} + (1 - \pi_{\ell j}^{(1)})/(r-1)$$

$$w' = (w_1, \cdots w_{r-1})$$

Examination of Eq. 6 shows that this is a system of $(r-1)$ equations each of which (in general) is of degree $2n(r-1) - 1$. Thus, if 100 subjects are interviewed on a question with, say, three possible responses, Eq. 6 represents a system of two equations, each of which is of degree 399.

Implicit in Eq. 6 was the assumption that there existed θ_k's maximizing Eq. 5b, which lie in the unit interval. When such θ_k's do not exist, the alternative to treating the problem with a programming algorithm is to select the value zero or one, whichever yields the larger L^*. This remark applies also to the special cases considered below.

MULTICATEGORY RESPONSES WITH INDEPENDENCE

If X_{kj} and Y_j are independent random variables, each subject lies independently of the category he occupies. Then, from Eq. 1b and Eq. 2a,

$$\pi_{kj}^{(0)} = P\{Y_j = 1\} = \pi_{kj}^{(1)} \equiv \pi_j \tag{7}$$

Under these conditions the constraint of Eq. 2b is trivial, so that the likelihood function given in Eq. 5a may be used with Eq. 7 instead of that given in Eq. 5b. Without concern for whether the θ_k lie in the unit interval, or sum to one, conventional differentiation shows that the θ_k satisfy the equation system

$$\sum_{j=1}^{n} \frac{z_{kj}\left(\pi_j - \frac{1 - \pi_j}{r - 1}\right)}{\theta_k \pi_j + (1 - \theta_k)\frac{(1 - \pi_j)}{r - 1}} = \sum_{j=1}^{n} \frac{z_{rj}\left(\pi_j - \frac{1 - \pi_j}{r - 1}\right)}{1 - \sum_{m-1}^{r-1}\left[\theta_m \pi_j + (1 - \theta_m)\frac{(1 - \pi_j)}{r - 1}\right]} \tag{8}$$

for $(r\pi_j - 1)/(r - 1) \neq 0$ for some j. Now each equation in the system is of degree $(2n - 1)$, as will also be true for the remaining cases.

Next examine the likelihood function in Eq. 5a from which these equations were derived. The concavity argument following Eq. 5b is

seen to be applicable again. Hence, $L^*(\theta_1, \ldots, \theta_{r-1})$ is a concave function and can be maximized for $\theta \equiv (\theta_1, \ldots, \theta_{r-1})$ in the simplex included between the axes and a hyperplane passing through all unit vectors at their terminals by conventional convex programming algorithms. The solution will, of course, have the desirable property that any feasible local maximum found will also be a global maximum (this was also the case in Sec. II).

TWO-CATEGORY RESPONSE CASE

Since many surveys involve only questions with two possible responses, this special case is of particular interest and is therefore evaluated separately. Letting r = 2 in Eq. 6 shows that in this case, the maximum likelihood estimators satisfy

$$\sum_{j=1}^{n} \frac{z_{1j}(\pi_{1j}^{(1)} + \pi_{2j}^{(1)} - 1)}{\theta_1(\pi_{1j}^{(1)} + \pi_{2j}^{(1)} - 1) + (1 - \pi_{2j}^{(1)})} = \sum_{j=1}^{n} \frac{z_{2j}(\pi_{1j}^{(1)} + \pi_{2j}^{(1)} - 1)}{\theta_1(1 - \pi_{1j}^{(1)} + \pi_{2j}^{(1)}) - \pi_{2j}^{(1)}} \tag{9}$$

or, if $|\theta_k| > 1$ in this system, the solution is zero or one.

Two-category Response Case with Independence

Setting k = 1 in Eq. 7 gives

$$\pi_{1j}^{(0)} = P\{Y_j = 1\} = \pi_{1j}^{(1)} = \pi_j$$

Then, the (unconstrained) maximum likelihood estimator of θ_1 satisfies (from Eq. 7)

$$\sum_{j=1}^{n} \frac{z_{1j}(2\pi_j - 1)}{\theta_1(2\pi_j - 1) + (1 - \pi_j)} = \sum_{j=1}^{n} \frac{z_{2j}(2\pi_j - 1)}{\theta_1(1 - 2\pi_j) + \pi_j} \tag{10}$$

III. ESTIMATOR EFFICIENCY AND NUMERICAL EXAMPLES

This Section develops the information matrix for this problem, for the case of independence. Since maximum likelihood estimators are asymptotically efficient and normally distributed (under mild regularity conditions, it is of interest to determine the asymptotic covariance matrix (the inverse of the information matrix).

INFORMATION MATRIX

The lower bounds for the variances of any estimators for this problem require the information matrix, J. Let $J \equiv (J_{km})$, for k, $m = 1, \ldots, r - 1$. Then

$$J_{km} = E\left(\frac{\partial L^*}{\partial \theta_m} \cdot \frac{\partial L^*}{\partial \theta_k}\right) \tag{11}$$

where L^* is defined in Eq. 5a, for the case of independence, by taking $\pi_{kj}^{(1)} = \pi_{kj}^{(0)} = \pi_j$.

Note that L^* may be expressed in the form

$$L^* = \sum_{j=1}^{n} \left\{ \sum_{m=1}^{r-1} Z_{mj} \ \ell n \ E(Z_{mj}) + Z_{rj} \ \ell n \left[1 - \sum_{m=1}^{r-1} E(Z_{mj})\right] \right\}$$

and if we define

$$v_j \equiv (\pi_j r - 1)/(r - 1), \text{ for } j = 1, \ldots, n,$$

$$\frac{\partial}{\partial \theta_k} E(Z_{kj}) = v_j \tag{12}$$

Hence,

$$\frac{\partial L^*}{\partial \theta_k} = \sum_{j=1}^{n} \left\{ \frac{Z_{kj}}{E(Z_{kj})} - \frac{Z_{rj}}{E(Z_{rj})} \right\} (v_j) \tag{13}$$

Taking expectations in Eq. 13 gives

$$E\left(\frac{\partial L^*}{\partial \theta_k}\right) = 0$$

Therefore, the minimum variance bounds for this problem exist, and may be computed by evaluating Eq. 11.

Accordingly, by direct algebraic computation, it can be found that if δ_{km} denotes the Kronecker delta,

$$J_{km} = \frac{1}{r-1} \sum_{j=1}^{n} (\pi_j r - 1)^2 \left\{ \frac{\delta_{km}}{p_k(\pi_j r - 1) + (1 - \pi_j)} + \frac{1}{p_r(\pi_j r - 1) + (1 - \pi_j)} \right\} \tag{14}$$

The diagonal elements of J^{-1} are, of course, the Cramér-Rao lower bounds for the variances of any unbiased estimators of the p_k's.

When $r = 2$, J reduces to a single element and substitution in Eq. 14 shows that for any unbiased estimator $\hat{p}_1$ of p_1,

$$\mathrm{Var}(\hat{p}_1) \geq \left\{ \sum_{j=1}^{n} \frac{(2\pi_j - 1)^2}{[p_1(2\pi_j - 1) + (1 - \pi_j)][\pi_j - p_1(2\pi_j - 1)]} \right\}^{-1} \tag{15}$$

and the same lower bound applies to $\mathrm{Var}(\hat{p}_2)$. Note that if $\pi_j = \frac{1}{2}$ for all j, Eq. 15 demands that $\hat{p}_1$ have infinite variance. However, by restricting $\hat{p}_1$ to the unit interval, $\mathrm{Var}(\hat{p}_1)$ will also be restricted

to the unit interval, so that a better bound in this case is unity. Note also that Eq. 5a shows that L^* is then not dependent upon θ_k. This case is discussed in Sec. VI.

EFFICIENCY

Since $L(z_1, \ldots, z_n|\theta_1, \ldots, \theta_{r-1})$ is a regular function of $\theta_1, \ldots, \theta_{r-1}$, i.e., $E(\partial L^*/\partial\theta_k) = 0$ for all k, the minimum variance bounds (MVB) for the parameters always exist. However, in general, the bounds cannot be attained (although, as will be seen, they are attainable in some cases). A necessary condition for attainability of the MVB, for all values of the parameters, is the existence of a sufficient statistic for the problem. But it is easy to check that for this problem one does not exist in general. Alternatively, one might attempt to find the Bhattacharyya bounds. This approach does not appear to be fruitful.

However, the existence of the MVB does provide the usual standard for measuring efficiency. Thus, if e_k denotes the efficiency of the estimator $\hat{\theta}_k$,

$$e_k = \frac{J^{-1}_{kk}}{\mathrm{Var}(\hat{\theta}_k)} \qquad k = 1, \ldots, r-1$$

and if $e_k = 1$, $\hat{\theta}_k$ is called efficient.

One special check case arises, for example, when $\pi_j = \alpha = \text{constant}$, for all j, and r = 2. In that case, the maximum likelihood estimates actually attain the MVB, uniformly in p_1 (or p_2) even for finite n. That is, for any fixed sample size, the variances of the estimators are at least as small as those of any other estimators (see Example 1

below). The fact that the MVB are attainable for this case follows immediately from the fact that in that special situation, there is a sufficient statistic, made up of the totals of people who claim each of the separate categories.

NUMERICAL EXAMPLES

Example 1

Assume $r = 2$ and that $\pi_{kj}^{(1)} = \pi_{kj}^{(0)} = \pi_j = \alpha$, where α is a constant for all j, k. When $\pi_j = \alpha$, L^* can be minimized without regard for the inequality constraints, which are then inactive. Substitution in Eq. 10 gives (recall that $p_1 = \theta_1$)

$$\hat{p}_1 = \frac{1}{(2\alpha - 1)n} \sum_{j=1}^{n} (z_{1j} + \alpha - 1)$$

and $\hat{p}_2 = 1 - \hat{p}_1$. Thus, if all subjects are truthful, $\alpha = 1$ and

$$\hat{p}_2 = \frac{1}{n} \sum_{j=1}^{n} z_{2j} = \overline{z_2}$$

as expected. Conversely, if all subjects lie, $\alpha = 0$ and

$$\hat{p}_2 = \frac{1}{n} \sum_{j=1}^{n} (1 - z_{2j}) = 1 - \overline{z_2}$$

The variances of these estimators are easily found. From the estimation result for $\hat{p}_1$, above,

$$\mathrm{Var}(\hat{p}_1) = \frac{1}{(2\alpha - 1)^2 n^2} \mathrm{Var} \sum_{j=1}^{n} (z_{1j} + \alpha - 1) = \frac{1}{(2\alpha - 1)^2 n} \mathrm{Var}(z_{1j})$$

But because $P\{Z_{1j} = 1\} = p_1(2\alpha - 1) + (1 - \alpha)$ for this case,

$$\mathrm{Var}(Z_{1j}) = [p_1(2\alpha - 1) + 1 - \alpha][\alpha - p_1(2\alpha - 1)]$$

Substitution gives

$$\mathrm{Var}(\hat{p}_1) = \frac{[p_1(2\alpha - 1) + (1 - \alpha)][\alpha - p_1(2\alpha - 1)]}{(2\alpha - 1)^2 n}$$

Evaluation of Eq. 15 for $\pi_j = \alpha$ shows that the MVB is identical with the result just obtained for the maximum likelihood estimator, showing the latter is efficient in finite samples.

Example 2

Take $r = 2$, and $n = 2m$, $m = 1, 2, \ldots$. Assume, moreover, that

$$\pi_{kj}^{(1)} = \pi_{kj}^{(0)} = \pi_j = \begin{cases} \alpha_1 & j = 1, \ldots, m \\ \alpha_2 & j = m + 1, \ldots, 2m \end{cases}$$

That is, half of the subjects lie with probability $(1 - \alpha_1)$ and the other half lie with probability $(1 - \alpha_2)$. For simplicity, take $\alpha_1 = 1$ and $\alpha_2 = 0$. Then from Eq. 10,

$$\hat{p}_1 = \frac{\sum_{j=1}^{m} z_{1j} + \sum_{j=m+1}^{2m} (1 - z_{1j})}{n}$$

Since

$$\mathrm{Var}(\hat{p}_1) = \frac{1}{n^2}\left[\mathrm{Var}\sum_{1}^{m} z_{1j}\right] + \frac{1}{n^2}\left[\mathrm{Var}\sum_{m+1}^{2m} (1 - z_{1j}\right]$$

and since $P\{Z_{1j} = 1\} = p_1(2\alpha - 1) + (1 - \alpha)$ when $\pi_j = \alpha$,

$$\mathrm{Var}(z_{1j}) = \begin{cases} p_1(1 - p_1) & j = 1, \ldots, m \\ (1 - p_1)p_1 & j = m + 1, \ldots, 2m \end{cases}$$

Hence,

$$\mathrm{Var}(\hat{p}_1) = \frac{p_1(1 - p_1)}{n}$$

Substitution into Eq. 15 to evaluate the MVB gives

$$\mathrm{Var}(\hat{p}_1) \geq \left\{ \frac{m(2\alpha_1 - 1)^2}{[p_1(2\alpha_1 - 1) + 1 - \alpha_1][\alpha_1 - p_1(2\alpha_1 - 1)]} + \frac{m(2\alpha_2 - 1)^2}{[p_1(2\alpha_2 - 1) + (1 - \alpha_2)][\alpha_2 - p_1(2\alpha_2 - 1)]} \right\}^{-1}$$

or,

$$\mathrm{Var}\ \hat{p}_1 \geq \frac{p_1(1 - p_1)}{2m} = \frac{p_1 q_1}{n}$$

Hence, again in this case the maximum likelihood estimators are efficient in finite samples.

IV. ASSESSMENT OF MISCLASSIFICATION PROBABILITIES AND RELATED EFFECTS

The above estimation procedures have been developed on the basis that the probabilities that the subject's responses coincide with the true situation are known or may in some way be determined. Indeed, there are many situations in which the $\pi_{kj}^{(1)}$'s may be determined on the basis of collateral information. This section considers the techniques for assessing the misclassification probabilities and the effect of ignoring hostile subjects in the analysis.

ASSESSMENT TECHNIQUES

The technique that should be used to assign these probabilities varies with the circumstances. Often it may be possible to decide upon the reliability of responses to certain questions on the basis of the subject's answers to other questions about which the interviewer has personal knowledge and additional information. In some surveys, the behavior of the subject during the interview might be the only available basis for a rational assessment, whereas in severe circumstances polygraph instruments or drugs might serve as the main bases for assessment. A quantitative measure that depends upon the length of the subject's response or the total time of the interview might also be incorporated into the decisionmaking associated with π_j. In the personality questionnaires often given to employees or prospective employees of a company, a certain subset of "test questions" usually serves to establish the reliability of the subject's responses. In general, every effort should be made in designing the questionnaire so that the difficulty of assessing the π_{kj}'s is minimized.

In summary, there are available many indicators of whether or not a subject is falsifying his responses to an interviewer, and they may be used individually or in combination to obtain estimates of the truth probabilities.

A BAYESIAN APPROACH

A direct justification for replacing the $\pi_{kj}^{(1)}$ parameters by their "best guess" estimates may be found in the Bayesian approach. Now, instead of assuming the $\pi_{kj}^{(1)}$ are known parameters (as assumed in Eq. 1b), assume they are random with known mean values, M_{kj}, so that $M_j' \equiv (M_{1j}, \ldots, M_{rj})$. Define the r-vectors:

$$a_m = \frac{1}{1 - r} p - \frac{r}{1 - r} p_m \mu_m$$

and let

$$b_m = \frac{1 - p_m}{r - 1}$$

where μ_m is the vector of zeros with a one in the mth place, and m - 1, ..., r. Then, it may be checked that Eq. 5b may be equivalently written (as the likelihood function expressed as a function of the $\pi^{(1)}$'s)

$$L\left(z_1, \ldots, z_n \middle| (\pi_1^{(1)}, \ldots, \pi_n^{(1)})\right) = \prod_{j=1}^{n} \prod_{m=1}^{r} \left(a_m' \pi_j^{(1)} + b_m\right)^{z_{mj}} \tag{16}$$

where $\left[\pi_j^{(1)}\right]' = \left[\pi_{1j}^{(1)}, \ldots, \pi_{rj}^{(1)}\right]$, and (a_m, b_m) are independent of $\pi_j^{(1)}$. Since z_{mj} may only take on the values zero and one, L may be written in the equivalent form

$$L = \prod_{j=1}^{n} \sum_{m=1}^{r} z_{mj}\left(a_m'\pi_j^{(1)} + b_m\right) \tag{17}$$

Integrating to find the marginal likelihood function gives

$$L(z_1, \ldots, z_n|p) = \int \prod_{j=1}^{n} \sum_{m=1}^{r} z_{mj}\left(a_m'\pi_j^{(1)} + b_m\right) f\left(\pi_1^{(1)}, \ldots, \pi_n^{(1)}\right) d\pi^{(1)}$$

where $f\left(\pi_1^{(1)}, \ldots, \pi_n^{(1)}\right)$ denotes the joint density of $\left(\pi_1^{(1)}, \ldots, \pi_n^{(1)}\right)$ Assume, for simplicity, that there is no collusion among subjects and take

$$f\left(\pi_1^{(1)}, \ldots, \pi_n^{(1)}\right) = \prod_{j=1}^{n} f_j\left(\pi_j^{(1)}\right)$$

where the $f_j\left(\pi_j^{(1)}\right)$ are the densities of the $\pi_j^{(1)}$ vectors. Then,

$$L(z_1, \ldots, z_n|p) = \prod_{j=1}^{n} \sum_{m=1}^{r} z_{mj} \int \left(a_m'\pi_j^{(1)} + b_m\right) f_j\left(\pi_j^{(1)}\right) d\pi_j^{(1)}$$

$$= \prod_{j=1}^{n} \sum_{m=1}^{r} z_{mj}\left(a_m'M_j + b_m\right)$$

Replacing z_{mj} into the exponential form yields the marginal likelihood function

$$L(z_1, \ldots, z_n|p) = \prod_{j=1}^{n} \prod_{m=1}^{r} \left(a_m'M_j + b_m\right)^{z_{mj}} \tag{18}$$

Note that Eq. 18 is the same function of "p" as is Eq. 16, with the $\pi_j^{(1)}$'s replaced by their expected values. Hence, the resulting

programming problem yields the same estimates of $(p_1, \ldots, p_r)$ if the $\pi_j^{(1)}$ are replaced by their expected values, which are certainly equivalents for this problem.

EFFECT OF IGNORING HOSTILE SUBJECTS

The question naturally arises of whether it is worthwhile using the type of analysis recommended in this Memorandum, i.e., is there much saving over just ignoring the effect of misclassified subjects? The answer is that the saving can be slight, or it can be so large as to make it mandatory to take some corrective action, depending upon the situation. The effect is illustrated quantitatively below for the case of two categories, and independence. The example used to evaluate this problem can be used also as an approximate estimation technique.

Suppose R percent of the subjects interviewed claim to belong to category one. Then, if the hostility effect is ignored, the usual estimator of p_1 (maximum likelihood) gives $100\ \hat{p}_1 = R$.

Next, suppose that a fraction, α, of the subjects claiming category one, lie, and that a fraction, β, of the subjects claiming category two, lie. Then, it is easy to see intuitively that an estimator which accounts for the liars is given by

$$100\ \hat{\hat{p}}_1 = R(1 - \alpha) + \beta(100 - R) \tag{19}$$

In fact, exactly this result is obtained from Eq. 10 by making substitutions

$$z_{1j} = \begin{cases} 1, & j = 1, \ldots, \frac{nR}{100} \\ 0, & j = \frac{nR}{100} + 1, \ldots, n \end{cases} \qquad z_{2j} = \begin{cases} 1 - z_{1j} \\ j = 1, \ldots, n \end{cases}$$

$$\pi_j = \begin{cases} 0, & j = 1, \ldots, \frac{\alpha nR}{100} \\ 1, & j = \frac{\alpha nR}{100} + 1, \ldots, \frac{nR}{100} \end{cases}$$

$$\pi_j = \begin{cases} 0, & j = \frac{nR}{100} + 1, \ldots, \frac{nR}{100} + \beta\left(n - \frac{nR}{100}\right) \\ 1, & j = \frac{nR}{100} + \beta\left(n - \frac{nR}{100}\right) + 1, \ldots, n \end{cases}$$

Now define ϵ^* to be the absolute error (expressed in percent probability) made by ignoring the effect of hostile subjects. Then from Eq. 19

$$\epsilon^* = \left|\hat{\hat{p}}_1 - \hat{p}_1\right| 100 = \left|\beta(100 - R) - R\alpha\right| \tag{20}$$

Examination of Eq. 20 shows that ϵ^* can be anything between 0 and 100 percent, depending upon the values of (α, β, R).

For example, if all subjects lie, $\alpha = 1$, $\beta = 1$. Then from Eq. 20, $\epsilon = |100 - 2R|$. Thus, if $R = 100$, ϵ is 100 percent, whereas if $R = 50$, $\epsilon = 0$. All varieties of intermediate results may be obtained by considering excursions of α, β, and R.

V. MULTIPLE QUESTION ANALYSIS

This Section is addressed to the problem of estimating the multinomial population proportions for many questions, simultaneously. This problem is more complex than the single question case for two reasons. One reason is that some subjects may exhibit inconsistent behavior, in that their truth telling probabilities may vary over question number. The second reason is that the responses of a given subject to many questions may be correlated. There will be a multivariate probability distribution generated by the joint probability of actually belonging to category k for question number i_1, and belonging to category k' for question number i_2, etc. The problem is clearly much more difficult, but we can examine what is involved.

Suppose consistent behavior can be assumed. Define $p_k^{(i_1)}$ as the probability of belonging to category k for question i_1, and let $Z_{kj}^{(i_1)}$ denote the value of Z_{kj} for question i_1. We require simultaneous estimation of the category probabilities for each question. In particular, it is desired to find estimators of the two-dimensional marginal probabilities, $p_{k,k'}^{(i_1,i_2)}$, of falling in category k on question i_1, and category k' on question i_2, and of the higher order marginal probabilities, which are more complicated. These probabilities can be developed by evaluating the covariance matrix of the jointly distributed $Z_{kj}^{(i)}$.

VI. A SPECIAL CASE AND AN OPPOSITION STRATEGY

In this Section it is shown that the one case in which the analysis cannot be used is the one corresponding to maximum entropy or information content in the set of responses. The sense of "information" used here is that of Shannon (see, for example, Khinchin[5]). Based upon this result, statistical inference on the p_k's is hopeless, although at the same time, motivation is provided for the development of an optimal strategy for hostile subjects.

Recall that the assumption following Eq. 8 required that

$$\frac{r\pi_j - 1}{r - 1} \neq 0$$

for at least one j. Clearly, unless this is true, Eq. 8 yields no information and the entire analysis (for the case of independence) breaks down. When r = 2 and independence applies, the assumption requires that

$$2\pi_j - 1 \neq 0$$

for at least one j; i.e., it is required that there exist at least one j for which $\pi_j \neq 1/2$.

The r events corresponding to the jth subject's response falling into the kth category, k = 1, ..., r, have associated probabilities (see Eq. 4, and take $\pi_{kj}^{(1)} = \pi_{kj}^{(0)} = \pi_j$)

$$P_{kj} \equiv P\{Z_{kj} = 1\} = p_k\left(\frac{\pi_j r - 1}{r - 1} + \frac{1 - \pi_j}{r - 1}\right)$$

Since these r events partition the space of possible events, the jth interview corresponds to an experiment whose information content is defined as

$$H_j \equiv - \sum_{k=1}^{r} P_{kj} \ln P_{kj}$$

It is widely known (and trivial to show) that H_j is maximized when $P_{kj} = 1/r$; that is, when

$$p_k\left(\frac{\pi_j r - 1}{r - 1}\right) + \frac{1 - \pi_j}{r - 1} = \frac{1}{r}$$

But this equation must hold identically in p_k. Therefore, it is necessary that

$$\frac{\pi_j r - 1}{r - 1} = 0, \quad \frac{1 - \pi_j}{r - 1} = \frac{1}{r}$$

Substitution shows that these equations require that

$$\pi_j = \frac{1}{r}$$

for all j. For the two-category independence case, $\pi_j = \frac{1}{2}$ for all j yields the maximum information. Thus, failure of the assumption required in the analysis corresponds to maximum entropy or disorder in the set of responses of the subjects.

If such a case arose in practice, maximum likelihood estimators could not be used. For this reason, it is clear that if hostile subjects were aiming at an optimal strategy, they would all lie independently of the categories they occupy, and would randomize their

responses. They would tell the truth 100 (1/r) percent of the time (when there are r possible responses to a question), and lie half the time when they are in dichotomous response situations.

REFERENCES

1. Warner, Stanley L., "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," Journal of the American Statistical Association, Vol. 60, 1965, pp. 63-69.

2. Hendricks, Walter A., "Estimation of the Probability that an Observation Will Fall into a Specified Class," Journal of the American Statistical Association, Vol. 59, 1964, pp. 225-232.

3. Mote, V. L., and R. L. Anderson, "An Investigation of the Effect of Misclassification on the Properties of χ-tests in the Analysis of Categorical Data," Biometrika, Vol. 52, 1965, pp. 95-109.

4. Rosen, J. B., "The Gradient Projection Method for Nonlinear Programming, Part I, Linear Constraints," Journal of the Society for Industrial and Applied Mathematics, Vol. 8, 1960, pp. 181-217.

5. Khinchin, A. I., "The Entropy Concept in Probability Theory," Uspekhi Matematicheskikh Nauk, Vol. 8, No. 3, 1953, pp. 3-20 (translated in Khinchin's Mathematical Foundations of Information Theory, New York: Dover Publications, Inc., 1957).

Estimating from Misclassified Data, S. J. Press, RM-5360-ISA/ARPA, July 1967

Conducted by: The RAND Corporation For: ISA and ARPA/AGILE

PURPOSE: To provide a statistical method for estimating the proportions of items in each of several categories, based on an item-by-item classification in which many items may be misclassified. A specific case of interest is that in which the items are subjects being interviewed and the subjects may give false responses.

RELATED TO: Motivation and morale studies of the Viet Cong conducted by RAND for ARPA and ISA.

METHODOLOGY AND DISCUSSION: In sampling from a human population it is usually assumed that the interviewee is cooperative and that his responses to questions correspond to the true situation as far as he is concerned. Thus, in a sample size n with S_n individuals who claim to have a given characteristic, $S_{n/n}$ is the maximum likelihood estimator of the population proportion corresponding to that characteristic. If some of the people questioned are hostile to the interviewer in that they give false responses for some reason, then $S_{n/n}$ is no longer a reasonable estimator, and a different procedure must be used.

FINDINGS: This study developed maximum likelihood estimators of the category proportions for both the two-category and the multicategory response cases with respect to a group of noncooperative interviewees. An assessment is made, for each subject, of the probability that he is hostile. These probabilities are then combined with the actual responses to yield the maximum likelihood estimators. Explicit evaluation of the estimators for a sample of n subjects and r categories requires solution of a simple concave programming problem involving a logarithmic objective function in variables confined to the unit interval. A bayesian approach is used to evaluate the misclassification (or hostility) probabilities. It is assumed that the analysis applies to a single question only. For a survey containing many questions the estimators would be evaluated separately for each question. Moreover, indication is given of how the analysis can be generalized to consider many questions simultaneously.

A special case in which the analysis cannot be used is the one corresponding to maximum entropy or information content in the set of responses. In this case, if hostile subjects were aiming at an optimal strategy, they would all lie independently of the categories they occupy and would randomize their responses: They would tell the truth 100(1/r) percent of the time (when there are r possible responses to a question), and lie half the time when they are in dichotomous response situations.

POTENTIAL FOR FURTHER DEVELOPMENT: Press has laid the theoretical groundwork for a reanalysis of many of the interviews already taken from Hoi Chanh and Tu Binh. Certainly some rigorous empirical testing of the theoretical estimates ought to be carried out to check the validity of Press' findings and hypotheses.

EVALUATION: Bayesian statistics has its prestigeous supporters and its equally influential detractors. One should imagine that a marriage of the empirical data from the Motivation and Morale study and Mr. Press' techniques would resolve the issue, at least in this instance. Until that occurs, one might only note that the approach is solid, the concepts are clear, and the analysis seems reasonable.

Prepared by: G D Brewer
G. D. Brewer

Approved by: _____
S. J. Deitchman

Estimating from Misclassified Data, S. J. Press, RM-5360-ISA/ARPA, July 1967

Conducted by: The RAND Corporation For: ISA and ARPA/AGILE

PURPOSE: To provide a statistical method for estimating the proportions of items in each of several categories, based on an item-by-item classification in which many items may be misclassified. A specific case of interest is that in which the items are subjects being interviewed and the subjects may give false responses.

RELATED TO: Motivation and morale studies of the Viet Cong conducted by RAND for ARPA and ISA.

METHODOLOGY AND DISCUSSION: In sampling from a human population it is usually assumed that the interviewee is cooperative and that his responses to questions correspond to the true situation as far as he is concerned. Thus, in a sample size n with S_n individuals who claim to have a given characteristic, $S_{n/n}$ is the maximum likelihood estimator of the population proportion corresponding to that characteristic. If some of the people questioned are hostile to the interviewer in that they give false responses for some reason, then $S_{n/n}$ is no longer a reasonable estimator, and a different procedure must be used.

FINDINGS: This study developed maximum likelihood estimators of the category proportions for both the two-category and the multicategory response cases with respect to a group of noncooperative interviewees. An assessment is made, for each subject, of the probability that he is hostile. These probabilities are then combined with the actual responses to yield the maximum likelihood estimators. Explicit evaluation of the estimators for a sample of n subjects and r categories requires solution of a simple concave programming problem involving a logarithmic objective function in variables confined to the unit interval. A bayesian approach is used to evaluate the misclassification (or hostility) probabilities. It is assumed that the analysis applies to a single question only. For a survey containing many questions the estimators would be evaluated separately for each question. Moreover, indication is given of how the analysis can be generalized to consider many questions simultaneously.

A special case in which the analysis cannot be used is the one corresponding to maximum entropy or information content in the set of responses. In this case, if hostile subjects were aiming at an optimal strategy, they would all lie independently of the categories they occupy and would randomize their responses: They would tell the truth 100(1/r) percent of the time (when there are r possible responses to a question), and lie half the time when they are in dichotomous response situations.

POTENTIAL FOR FURTHER DEVELOPMENT: Press has laid the theoretical groundwork for a reanalysis of many of the interviews already taken from Hoi Chanh and Tu Binh. Certainly some rigorous empirical testing of the theoretical estimates ought to be carried out to check the validity of Press' findings and hypotheses.

EVALUATION: Bayesian statistics has its prestigeous supporters and its equally influential detractors. One should imagine that a marriage of the empirical data from the Motivation and Morale study and Mr. Press' techniques would resolve the issue, at least in this instance. Until that occurs, one might only note that the approach is solid, the concepts are clear, and the analysis seems reasonable.

Prepared by: G. D. Brewer
G. D. Brewer

Approved by: _____
S. J. Deitchman

Estimating from Misclassified Data, S. J. Press, RM-5360-ISA/ARPA, July 1967

Conducted by: The RAND Corporation For: ISA and ARPA/AGILE

PURPOSE: To provide a statistical method for estimating the proportions of items in each of several categories, based on an item-by-item classification in which many items may be misclassified. A specific case of interest is that in which the items are subjects being interviewed and the subjects may give false responses.

RELATED TO: Motivation and morale studies of the Viet Cong conducted by RAND for ARPA and ISA.

METHODOLOGY AND DISCUSSION: In sampling from a human population it is usually assumed that the interviewee is cooperative and that his responses to questions correspond to the true situation as far as he is concerned. Thus, in a sample size n with S_n individuals who claim to have a given characteristic, $S_{n/n}$ is the maximum likelihood estimator of the population proportion corresponding to that characteristic. If some of the people questioned are hostile to the interviewer in that they give false responses for some reason, then $S_{n/n}$ is no longer a reasonable estimator, and a different procedure must be used.

FINDINGS: This study developed maximum likelihood estimators of the category proportions for both the two-category and the multicategory response cases with respect to a group of noncooperative interviewees. An assessment is made, for each subject, of the probability that he is hostile. These probabilities are then combined with the actual responses to yield the maximum likelihood estimators. Explicit evaluation of the estimators for a sample of n subjects and r categories requires solution of a simple concave programming problem involving a logarithmic objective function in variables confined to the unit interval. A bayesian approach is used to evaluate the misclassification (or hostility) probabilities. It is assumed that the analysis applies to a single question only. For a survey containing many questions the estimators would be evaluated separately for each question. Moreover, indication is given of how the analysis can be generalized to consider many questions simultaneously.

A special case in which the analysis cannot be used is the one corresponding to maximum entropy or information content in the set of responses. In this case, if hostile subjects were aiming at an optimal strategy, they would all lie independently of the categories they occupy and would randomize their responses: They would tell the truth 100(1/r) percent of the time (when there are r possible responses to a question), and lie half the time when they are in dichotomous response situations.

POTENTIAL FOR FURTHER DEVELOPMENT: Press has laid the theoretical groundwork for a reanalysis of many of the interviews already taken from Hoi Chanh and Tu Binh. Certainly some rigorous empirical testing of the theoretical estimates ought to be carried out to check the validity of Press' findings and hypotheses.

EVALUATION: Bayesian statistics has its prestigeous supporters and its equally influential detractors. One should imagine that a marriage of the empirical data from the Motivation and Morale study and Mr. Press' techniques would resolve the issue, at least in this instance. Until that occurs, one might only note that the approach is solid, the concepts are clear, and the analysis seems reasonable.

Prepared by: G D Brewer
G. D. Brewer

Approved by: _____
S. J. Deitchman