ESTIMATING A BERNULLI PARAMETER FROM A SAMPLE OF MISCLASSIFIED RESPONSES AND A SUB-SAMPLE OF RANDOMIZED RESPONSES

by

Yosef Hochberg

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1016

June 1975

Estimating a Bernulli Parameter from a Sample of Misclassified Responses and a Sub-Sample of Randomized Responses

by

Yosef Hochberg

University of North Carolina at Chapel Hill

ABSTRACT

It appears that in the various publications on the use of the Randomized Response technique it has always been assumed that the experimenter has available to him only the sample of Randomized Responses to draw inferences from.

However, in many applications, the Randomized Response technique is used when an original, usually large, sample is available. The original sample is based on misclassified responses due to some stigma in the issues under study.

In this note we assume that a sub-sample of individuals from the original sample (with the individual misclassified responses available) is taken for application of the Randomized Response technique.

Based on the simultaneous classification of the sub-sampled individuals according to their misclassified and randomized responses and the original total sample of misclassified responses, efficient methods for estimating the Bernulli parameter of a stigmatizing response are discussed.

1. Introduction

Tenenbein [1970] proposed a double sampling technique for the problem of estimating from categorical data which is subject to misclassification errors. His approach is based on the model in Bross [1954] for misclassification errors. The following experimental situation is assumed. There are two classification devices available. One device is expensive to apply and gives correct results. The other device is relatively inexpensive but fallible. Such experimental situations were considered by many other writers, e.g., Diamond and Lilienfeld [1962], where the true expensive classification device is a physician's examination whereas the fallible classifier is a patient's questionnaire. The double sampling scheme involves the following steps:

a. Obtain the fallible classifications on a "large" number of units -- N, say

b. Obtain in addition the true classifications on a sub-sample of n out of the N units

c. Combine (a) and (b) efficiently for estimating the Bernulli parameter under study.

Note that in many problems, stage (a) might simply amount to accessing an existing file.

However, many experimental situations do arise where there is no exact device for measuring a true response. Such are all experimental

situations when "practically" only the individual can know the true response. If the response has a stigmatizing nature, estimates based on a direct questionnaire are biased due to errors of misclassification. A reasonable approach in such cases would be to replace phase (b) above by a sub-sample of individuals on which the randomized response is used. (The randomized response proposed by Warner [1965] was extended and modified by many writers. For a summary see Greenberg et al [1974].)

Obviously, the two stage approach will be more efficient than using only the randomized response sub-sample when the sample of phase (a) is available. In cases it is not, the double sampling plan might still be considered a good strategy depending on obvious parameters. This is discussed in the following section.

2. A Double Sampling Plan with Misclassified and Randomized Responses

2.1 Introduction and Notation

The randomized response to be used in our discussion is the one that uses an unrelated question of a known proportion. This technique can always be used by artificially forming responses with known probabilities, see Greenberg et al [1974].

The data in the sub-sample can be summarized as in Table 1,

TABLE 1

FREQUENCIES IN THE SUB-SAMPLE

		Misclassified Responses		
		No	Yes	Total
Randomized Responses	No	N_{00}	N_{01}	N_{0}
	Yes	N_{10}	N_{11}	$N_{1.}$
	Total	$N_{.0}$	$N_{.1}$	N

A "Yes" and "No" for the misclassified responses means belonging or not to the stigmatizing group, respectively (as given by that response). The "Yes" and "No" for the randomized response are the literal responses when the individual is asked: "Do you belong to the stigmatizing group?" with probability p and with probability 1-p he is asked an unrelated question which has a probability w for a "Yes" response and 1-w for a "No" response. Both $0<p<1$ and $0<w<1$ are assumed known.

Corresponding to Table 1 there is the table of population proportions.

TABLE 2

POPULATION PROPORTIONS

		Misclassified Responses		
		No	Yes	Total
Randomized Responses	No	β_{00}	β_{01}	$1-\lambda$
	Yes	β_{10}	β_{11}	λ
	Total	$1-\phi$	ϕ	

On letting $\alpha_{0,1}$ denote the probability that the fallible classifier gives "No" when the truth is "Yes", letting $\alpha_{1,0}$ denote the reversed error and letting π denote the true proportion of individuals in the stigmatizing group, we have the following identities:

$$\lambda = p\pi + (1-p)w$$

$$\beta_{00} = (1-\pi)(1-\alpha_{1,0})[p+(1-p)(1-w)]+\pi\alpha_{0,1}(1-p)(1-w)$$

$$\beta_{01} = \pi(1-\alpha_{0,1})(1-p)(1-w) + (1-\pi)\alpha_{1,0}[p + (1-p)(1-w)]$$

$$\beta_{10} = (1-\pi)(1-\alpha_{1,0})(1-p)w + \pi\alpha_{0,1}[p + (1-p)w]$$

$$\beta_{11} = \pi(1-\alpha_{0,1})[p + (1-p)w] + (1-\pi)\alpha_{1,0}(1-p)w.$$

The N-n remaining individuals for which information is only available from one source can be presented in a table as follows:

	No	Yes	Total
Frequency	Y	X	N-n
Expected Proportion	$1-\phi$	ϕ	1

2.2 Maximum Likelihood Estimates of $\alpha_{0,1}$, $\alpha_{1,0}$, and π

First we introduce ξ and Ψ where,

$\xi = \beta_{11}/\phi$ = Conditional probability of getting a "Yes" for the randomized response when the fallible response is "Yes".

$\Psi = \beta_{10}/(1-\phi)$ = Conditional probability of getting a "Yes" for the randomized response when the fallible response is "No".

The common probability distribution function of X, Y, n_{00}, n_{10}, n_{01} and n_{11} is given by

$$(\text{constant})\phi^{X+n_{\cdot 1}}(1-\phi)^{Y+n_{\cdot 0}}\xi^{n_{11}}(1-\xi)^{n_{01}}\Psi^{n_{10}}(1-\Psi)^{n_{00}} .$$

The Maximum Likelihood Estimators (MLE) of ϕ, ξ, Ψ are given by $\hat{\phi}$, $\hat{\xi}$, $\hat{\Psi}$, respectively, where

$$\hat{\phi} = \frac{X+n_{\cdot 1}}{N} \quad ; \quad \hat{\xi} = \frac{n_{11}}{n_{\cdot 1}} \quad ; \quad \hat{\Psi} = \frac{n_{10}}{n_{\cdot 0}} \quad .$$

Now, since $(\phi, \xi, \Psi) \xleftrightarrow{1:1} (\lambda, \xi, \Psi)$ we get the MLE of λ via:

$$\lambda = \phi\xi + (1-\phi)\Psi ,$$

which in turn, from

$$\lambda = p\pi + (1-p)w$$

gives the MLE of π. We get

$$\hat{\lambda} = \frac{X+n_{\cdot 1}}{N}\frac{n_{11}}{n_{\cdot 1}} + \frac{Y+n_{\cdot 0}}{N}\frac{n_{10}}{n_{\cdot 0}}$$

$$\hat{\pi} = \frac{\hat{\lambda}-(1-p)w}{p} \quad .$$

To obtain the MLE's of $\alpha_{0,1}$ and $\alpha_{1,0}$ we write

$$\xi = \frac{\pi(1-\alpha_{0,1})[p+(1-p)w]+(1-\pi)\alpha_{1,0}(1-p)w}{\phi}$$

$$\Psi = \frac{(1-\pi)(1-\alpha_{1,0})(1-p)w+\pi\alpha_{0,1}[p+(1-p)w]}{1-\phi} \quad .$$

On letting $\underline{\theta} = (\xi,\Psi)'$,

$$\underline{b} = \begin{bmatrix} \dfrac{\pi[p+(1-p)w]}{\phi} \\ \dfrac{(1-\pi)(1-p)w}{1-\phi} \end{bmatrix} \quad , \quad \underline{A} = \begin{bmatrix} \dfrac{-\pi[p+(1-p)w]}{\phi} & \dfrac{(1-\pi)(1-p)(1-w)}{\phi} \\ \dfrac{\pi[p+(1-p)w]}{1-\phi} & \dfrac{-(1-\pi)(1-p)w}{1-\phi} \end{bmatrix}$$

and $\underline{\alpha}' = (\alpha_{0,1}\ \alpha_{1,0})$ we can write

$$\underline{\alpha} = \underline{A}^{-1}(\underline{\theta}-\underline{b}).$$

It is easily verified that $\underline{A}$ is non-singular provided $w \neq \frac{1}{2}$.

2.3 The Asymptotic Variance of $\hat{\pi}$

The asymptotic Variance-Covariance matrix of $(\hat{\phi},\hat{\xi},\hat{\Psi})$ is the inverse of the corresponding information matrix. It can be verified that asymptotically $\hat{\phi},\hat{\xi}$,$\hat{\Psi}$ are independent and

$$\mathrm{Var}(\hat{\phi}) = \frac{\phi(1-\phi)}{N}$$

$$\mathrm{Var}(\hat{\xi}) = \frac{\xi(1-\xi)}{n\phi}$$

$$\text{Var}(\hat{\Psi}) = \frac{\Psi(1-\Psi)}{n(1-\phi)} \quad .$$

Since $\hat{\lambda} = \hat{\xi}\hat{\phi} + \hat{\Psi}(1-\hat{\phi})$, we can obtain the asymptotic variance of $\hat{\lambda}$ by linearization which gives

$$V(\hat{\lambda}) = \text{Asymptotic} \quad \text{Var}(\hat{\lambda}) = (\xi-\Psi)^2 \frac{\phi(1-\phi)}{N} + \frac{\phi\xi(1-\xi)}{n} + \frac{(1-\phi)\Psi(1-\Psi)}{n} \quad .$$

From this expression and the relation between $\hat{\lambda}$ and $\hat{\pi}$ the asymptotic variance of $\hat{\pi}$ is obtained.

Next we express the variance of $\hat{\lambda}$ in terms of the $\alpha_{0,1}$ and $\alpha_{1,0}$. First we let

$\theta_{0,1}$ = Conditional probability of getting a "No" response on the fallible classifier when the randomized response is "Yes."

$\theta_{1,0}$ = Conditional probability of getting a "Yes" response on the fallible classifier when the randomized response is "No."

It then follows, as in Tenenbein [1970], that

$$V(\hat{\lambda}) = \frac{\lambda(1-\lambda)}{n}\left[1 - \frac{\lambda(1-\lambda)}{\phi(1-\phi)}(1-\theta_{0,1}-\theta_{1,0})^2\right] + \frac{[\lambda(1-\lambda)]^2}{N\phi(1-\phi)}(1-\theta_{0,1}-\theta_{1,0})^2 \quad .$$

Straightforward computations give

$$\theta_{0,1} + \theta_{1,0} = \frac{p\pi(1-\pi)(\alpha_{0,1}+\alpha_{1,0}) + (\lambda-\pi p)(1-\lambda) + p\pi(\pi-\lambda)}{\lambda(1-\lambda)} \quad .$$

This is obtained from any two of the four identities involving the β_{ij}'s, i,j = 0,1, in Section 2.1.

We get

$$1-\theta_{0,1}-\theta_{1,0} = p\frac{\pi(1-\pi)}{\lambda(1-\lambda)}(1-\alpha_{0,1}-\alpha_{1,0})$$

which can now be substituted in the expression for $V(\hat{\lambda})$ to give

$$V(\hat{\lambda}) = \frac{\lambda(1-\lambda)}{n}\left[1 - p^2 \frac{\pi(1-\pi)}{\phi(1-\phi)} \frac{\pi(1-\pi)}{\lambda(1-\lambda)} (1-\alpha_{0,1}-\alpha_{1,0})^2\right]$$

$$+ p^2 \frac{\pi(1-\pi)}{n} \frac{\pi(1-\pi)}{\phi(1-\phi)} (1-\alpha_{0,1}-\alpha_{1,0})^2 .$$

It is interesting to note the following:

(i) If $p = 1$ (i.e. $\lambda = \pi$) this case reduces to that of Tenenbein [1970].

(ii) If $0<p<1$ and no error is involved in using the fallible classifier, one does not get $V(\hat{\lambda}) = p^2\pi(1-\pi)/N$, because in such a case $\hat{\lambda}$ is not the MLE of λ. In this case ξ and Ψ are fixed constants

$$\xi = p + (1-p)w$$

$$\Psi = (1-p)w.$$

Clearly the MLE for π in this case is that of ϕ with variance $\pi(1-\pi)/N$.

2.4 The Least Squares Approach

The estimators considered so far were MLE's, similar to those in Tenenbein [1970]. We now consider least squares estimators (LSE) based on Grizzle et al [1969]. These are obtained along the line of Koch et al [1972]. The strategy is to obtain LSE's of the β_{ij}'s from which the estimators of λ, $\alpha_{0,1}$ and $\alpha_{1,0}$ are obtained.

Let $\underline{n}' = (n_{00}, n_{01}, n_{10}, n_{11})$, $\underline{N}' = (Y,X)$, $\underline{p} = (\underline{n}/n)$, $\underline{p}_N = (\underline{N}/(N-n))$, $\underline{p}_G' = (\underline{p}', \underline{p}_N')$, then

$$E(\underline{p}_G') = (\underline{\pi}', \underline{\pi}!) \text{ where } \underline{\pi}' = (\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}) \text{ and } \underline{\pi}! = (1-\phi, \phi).$$

The covariance matrix of $\underline{p}_G$ is given by

$$\underline{V}(\underline{\pi}) = \begin{bmatrix} (D_{\underline{\pi}} - \underline{\pi}\underline{\pi}')/n & \underline{0}:4\times 2 \\ \underline{0}:2\times 4 & (D_{\underline{\pi}_\cdot} - \underline{\pi}_\cdot\underline{\pi}_\cdot')/(N-n) \end{bmatrix},$$

where $D_{\underline{a}}$ is diagonal with $\underline{a}$ as the elements on the diagonal. Let

$$\underline{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 \\ \text{(Symmetric)} & & & & 0 & 0 \\ & & & & & 1 \end{bmatrix}$$

and write $\underline{F} = \underline{A}\underline{p}_G$. Next we let our model be $E(\underline{F}) = \underline{X}\underline{\beta}$, where

$$\underset{4\times 3}{\underline{X}} = \begin{bmatrix} \underline{I}:3\times 3 \\ \hline 0 \quad 1 \quad 1 \end{bmatrix}$$

and $\underline{\beta} = (\beta_{01}, \beta_{10}, \beta_{11})$. The LSE of $\underline{\beta}$ is $\underline{b}$

$$\underline{b} = (\underline{X}'\underline{V}_F^{-1}\underline{X})^{-1}\underline{X}'\underline{V}_F^{-1}\underline{F}$$

where $\underline{V}_F = \underline{A}'\underline{V}(\underline{p}_G)\underline{A}$. The resulting estimated variance-covariance of $\underline{b}$ is

$$\underline{V}_b = (\underline{X}'\underline{V}_F^{-1}\underline{X})^{-1}.$$

Having obtained the estimates of the β_{ij}'s (via $\hat{\beta}_{00}=1-b_1-b_2-b_3$, $\hat{\beta}_{01}=b_1$, $\hat{\beta}_{10}=b_2$, $\hat{\beta}_{11}=b_3$) one obtains $\hat{\lambda}=\hat{\beta}_{10}+\hat{\beta}_{11}$, $\hat{\phi}=\hat{\beta}_{01}+\hat{\beta}_{11}$ and goes on to obtain $\hat{\pi}$, $\hat{\alpha}_{0,1}$ and $\hat{\alpha}_{1,0}$ as done earlier with the MLE's.

2.5 The Coefficient of Reliability and Efficiency of the Two-Stage Procedure

The discussion here is based on the MLE in order to follow the line of Tenenbein [1970]. Let k_R denote the squared correlation coefficient between the randomized and the fallible responses (when 1 and 0 are attached to "Yes" and "No" respectively).

We have

$$k_R = \frac{\lambda(1-\lambda)}{\phi\gamma(1-\phi)} \quad (1-\theta_{1,0}-\theta_{0,1})^2.$$

On letting k_T denote the squared correlation between the true and the fallible responses we have

$$k_T = \frac{\pi(1-\pi)}{\phi(1-\phi)} \ (1-\alpha_{1,0}-\alpha_{0,1})^2.$$

As in Tenenbein [1970] we get here

$$V(\hat{\pi}) = \frac{\lambda(1-\lambda)}{p^2} \left\{ \frac{1}{n} [1-k_R] + \frac{1}{N} k_R \right\} \quad .$$

If only the randomized responses are utilized one gets

$$V(\hat{\hat{\pi}}) = \frac{\lambda(1-\lambda)}{np^2}$$

where $\hat{\hat{\pi}}$ is the so obtained estimator of π. To study the efficiency of the two stage procedure versus the use of only the randomized responses for equal cost, one must first obtain the best allocation of observations, i.e., that which achieves minimum $V(\hat{\pi})$ for a given cost. This is equivalent to minimizing $V(\hat{\lambda})$ for a given cost which follows along the line of Tenenbein [1970]. Thus, on letting c_2 and c_1 be the costs per unit sampling of a misclassified response, and of a randomized response respectively, letting $R = c_1/c_2$, $f = n/N$ and letting

$$f_0 = \text{Min} \left\{ \left(\frac{1-k_R}{k_R R} \right)^{\frac{1}{2}} \ , \ 1 \right\}$$

we have the best allocation

$$n = n_0 \left[\frac{Rf_0}{Rf_0+1}\right] \quad ; \quad N = (n_0-n)R$$

where $n_0 = C_0/c_1$ and C_0 is the total available cost.

The comparison of the two methods can thus be compared along the line of Tenenbein [1970] and we do not repeat this here.

Similar remarks can be made on a three stage type sampling as in Tenenbein [1971], where a pilot sample is taken to estimate k_R which then determines the best allocation.

Note that since $n<N$, for any $k_R>0$ the double sampling plan is more efficient when cost is not considered, (as is approximately the case when the total sample of misclassified responses are already available on file). It is interesting to compare the efficiency of $V(\hat{\pi})$ with that of $V(\hat{\hat{\pi}})$ for that important case.

We have

$$e = \frac{V(\hat{\pi})}{V(\hat{\hat{\pi}})} = 1-k_R(1-\frac{n}{N})$$

Thus, the efficiency (e) depends on k_R and n/N. However, k_R is clearly a function of p, w, $\alpha_{0,1}$, $\alpha_{1,0}$ and π.

In any particular problem the possibility of biased estimates from randomized responses should be considered too. There are too many parameters, thus tables are not considered here. However, in any specific problem the experimenter should make a decision, based on an appropriate pilot study, as to which course of action should be taken:

(i) Use only the sampling of misclassified responses.

(ii) Use only the randomized responses.

(iii) Use the double sampling scheme.

Such a decision should be based on one's best guesses of the relative costs, of π, of the errors $\alpha_{0,1}$, $\alpha_{1,0}$, the bias in the randomized response estimators and the parameters p and w.

REFERENCES

Bross, I. [1954]. Misclassification in 2×2 tables. Biometrics 10, 478-486.

Diamond, E. and Lilienfeld, A. [1962]. Effects of errors in classification and diagnosis in various types of epidemiological studies. American Journal of Public Health 10, 2106-2110.

Greenberg, B. G., Horvitz, D. G., and Abernathy, J. R. [1974]. A comparison of randomized response designs. Reliability and Biometry, Siam, Philadelphia. pp.787-815.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. [1969]. Analysis of categorical data by linear models, Biometrics 25, 489-504.

Koch, G. G., Imrey, P. B., and Reinfurt, D. W. [1972]. Linear model analysis of categorical data with incomplete response vectors. Biometrics 28, 663-692.

Tenenbein, A. [1970]. A double sampling scheme for estimating from binomial data with misclassifications. J. Amer. Statist. Assoc. 65, 1350-1361.

Tenenbein, A. [1971]. A double sampling scheme for estimating from binomial data with misclassifications of sample size determination. Biometrics 27, 935-944.

Warner, S. L. [1965]. Randomized response: a survey technique for eliminating evasive answer bias. J. Amer. Statist. Assoc. 60, 66-69.