COMBINING TWO DICHOTOMOUS RESPONSES CONTAINING RESPONSE ERRORS

Her Tzai Huang, Grambling College
Wayne A. Fuller, Iowa State University

Introduction

In sample surveys, response errors often constitute a sizeable portion of the total error associated with an estimator. Hansen, et al. [7] presented a mathematical model for survey observations containing response errors. Response errors in a binomial population were studied by Hansen, et al. [6]. They emphasized the difference in the impact of uncorrelated and correlated response deviations on the sampling properties of estimators. Since then a number of papers have been devoted to the effects of misclassification on estimates and tests associated with multinomial problems (e.g. [1], [5], [10], [12]).

Given the presence of response errors a questionnaire containing two responses for the variable of interest is sometimes used. For example, "How old are you?" and "When were you born?" can be used to obtain two (possibly different) responses for the age of sample individuals. The objective of such a questionnaire is to obtain a value for each individual somehow superior to which can be obtained from a single question. The presence of two questions requires a rule for combining the two answers.

There are a number of references dealing with the problem of combining estimators of a common mean if the sampling is from a normal distribution. If we have a number of estimates $y_i (i = 1,2,\ldots,k)$ normally and independently distributed about the same mean, μ, with known variances σ_i^2, the minimum variance unbiased estimator of μ is the weighted mean, $\bar{y}_w = \sum_{i=1}^{k} W_i y_i / W$, where $W_i = \sigma_i^{-2}$ and $W = \Sigma W_i$. When the σ_i^2 are unknown, they may be replaced by their unbiased estimators, s_i^2. Properties of such estimators have been investigated by Cochran [2], Meier [11], Cochran and Carroll [3], Huang [9] and others.

In this paper, we consider a finite universe in which individuals are classified into one of two classes, "1" or "0." The proportion of individuals in class 1 is denoted by P, and the proportion of individuals in class 0 denoted by Q, where $P + Q = 1$. We assume that a simple random sample of size n is drawn from this universe, and each individual responds to two questions that permit him to be classified into one of the two groups. These may be two questions on a single questionnaire or they could be questions on two different questionnaires. Due to the response errors, the two responses are not always the same. We consider the estimation of the population proportion P and the classification of sample individuals into the two classes. It is assumed that a super-population of responses exists for each individual. Let

- p_{mu} denote the probability that an individual who belongs to class 1 answers 1 to the m-th question;
- q_{mu} denote the probability that an individual who belongs to class 1 answers 0 to the m-th question;
- p_{mv} denote the probability that an individual who belongs to class 0 answers 1 to the m-th question;
- q_{mv} denote the probability that an individual who belongs to class 0 answers 0 to the m-th question.

We denote the response to the m-th question by sample individual i by Y_{mi} and assume that the response probabilities are such that the sample responses are unbiased for the population proportions, i.e.,

$$E(Y_{mi}) = p_{mu}P + p_{mv}Q = P, \qquad (1)$$

where the expectation is over individuals and responses.

Classification of Individuals Given a Third 0-1 Variable

In the continuous variable case it is possible to use the sample information to estimate weights by which the two responses may be combined (see Huang [9]). However, in the classification case additional information beyond that contained in the two responses seems to be required for efficient combination. We assume that a third zero-one variable, X_3, is available from the questionnaire. We also assume i) X_3 has nonzero correlation with the individual true value and ii) the response errors in Y are independent of X_3. Note that X_3 may contain response error provided that the response error is independent of that in Y.

Each individual response can be identified by one of eight 3-tuples, $Z_i = (Y_{1i}, Y_{2i}, X_{3i}) =$ (0, 0, 0), (1, 0, 0), (0, 1, 0), (1, 1, 0), (0, 0, 1), (1, 0, 1), (0, 1, 1), (1, 1, 1). We identify these eight possibilities by a single subscript, $j = 1,2,\ldots,8$ and let

- R denote the probability that X_{3i} equals 1;
- α denote the conditional probability that the true value for individual i, $\mu_{.i}$ equals 1 given that X_{3i} is 1;
- β denote the conditional probability that $\mu_{.i}$ equals 1 given that X_{3i} is 0;
- P_j denote the conditional probability that $\mu_{.i}$ equals 1 given case j, $j = 1,2,\ldots,8$;

$P_{(j)}$ denote the j-th conditional probability in the ordered arrangement of the P_j's,

$$P_{(1)} \le P_{(2)} \le \cdots \le P_{(8)},\ j = 1,2,\ldots,8.$$

$$P_{j_1j_2j_3} = P(Y_{1i} = j_1,\ Y_{2i} = j_2,\ X_{3i} = j_3),$$
$$j_1 = 0,1;\ j_2 = 0,1;\ j_3 = 0,1;$$

$F_{(j)} = P_{j_1j_2j_3}$ where the $F_{(j)}$ are ordered by the magnitude of P_j, $j = 1,2,\ldots,8$.

It is easily seen that $\beta(1-R) + \alpha R = P$ and the probabilities $P_{j_1j_2j_3}$ are given by

$$P_{000} = q_{1v}q_{2v}(1-\beta)(1-R) + q_{1u}q_{2u}\beta(1-R) \quad (2.1)$$

$$P_{100} = p_{1v}q_{2v}(1-\beta)(1-R) + p_{1u}q_{2u}\beta(1-R) \quad (2.2)$$

$$P_{010} = q_{1v}p_{2v}(1-\beta)(1-R) + q_{1u}p_{2u}\beta(1-R) \quad (2.3)$$

$$P_{110} = p_{1v}p_{2v}(1-\beta)(1-R) + p_{1u}p_{2u}\beta(1-R) \quad (2.4)$$

$$P_{001} = q_{1v}q_{2v}(1-\alpha)R + q_{1u}q_{2u}\alpha R \quad (2.5)$$

$$P_{101} = p_{1v}q_{2v}(1-\alpha)R + p_{1u}q_{2u}\alpha R \quad (2.6)$$

$$P_{011} = q_{1v}p_{2v}(1-\alpha)R + q_{1u}p_{2u}\alpha R \quad (2.7)$$

$$P_{111} = p_{1v}p_{2v}(1-\alpha)R + p_{1u}p_{2u}\alpha R \quad . \quad (2.8)$$

Further

$$P_1 = \frac{q_{1u}q_{2u}\beta}{q_{1v}q_{2v}(1-\beta) + q_{1u}q_{2u}\beta} \quad (3.1)$$

$$P_2 = \frac{p_{1u}q_{2u}\beta}{p_{1v}q_{2v}(1-\beta) + p_{1u}q_{2u}\beta} \quad (3.2)$$

$$P_3 = \frac{q_{1u}p_{2u}\beta}{q_{1v}p_{2v}(1-\beta) + q_{1u}p_{2u}\beta} \quad (3.3)$$

$$P_4 = \frac{p_{1u}p_{2u}\beta}{p_{1v}p_{2v}(1-\beta) + p_{1u}p_{2u}\beta} \quad (3.4)$$

$$P_5 = \frac{q_{1u}q_{2u}\alpha}{q_{1v}q_{2v}(1-\alpha) + q_{1u}q_{2u}\alpha} \quad (3.5)$$

$$P_6 = \frac{p_{1u}q_{2u}\alpha}{p_{1v}q_{2v}(1-\alpha) + p_{1u}q_{2u}\alpha} \quad (3.6)$$

$$P_7 = \frac{q_{1u}p_{2u}\alpha}{q_{1v}p_{2v}(1-\alpha) + q_{1u}p_{2u}\alpha} \quad (3.7)$$

$$P_8 = \frac{p_{1u}p_{2u}\alpha}{p_{1v}p_{2v}(1-\alpha) + p_{1u}p_{2u}\alpha} \quad (3.8)$$

We first develop a rule for classifying individuals assuming the population parameters known. We wish an assigned value for each individual, $\hat{\mu}_{.i}$, where $\hat{\mu}_{.i}$ is either zero or one and $\hat{\mu}_{.i}$ is chosen to minimize the expected squared error loss

$$E\{(\hat{\mu}_{.i} - \mu_{.i})^2\}$$

subject to the restriction that the mean of $\hat{\mu}_{.i}$ is an unbiased estimator of P,

$$E\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{.i}\} = P \quad . \quad (4)$$

We propose the following rule:

Rule: Order the P_j's such that

$$P_{(8)} \ge P_{(7)} \cdots \ge P_{(2)} \ge P_{(1)},$$

and define c to be the index such that

$$\sum_{j=c+1}^{8} F_{(j)} < P \text{ and } \sum_{j=c}^{8} F_{(j)} \ge P \quad .$$

Assign

$$\hat{\mu}_{.i(j)} = 1, \quad j=c+1, c+2, \ldots, 8$$

$$\hat{\mu}_{.i(c)} = 1, \text{ with probability A}$$

$$= 0, \text{ with probability 1-A}$$

$$\hat{\mu}_{.i(j)} = 0, \quad j=1,2,\ldots,c-1$$

where

$$A = \frac{1}{F_{(c)}}[P - F_{(8)} - F_{(7)} - \cdots - F_{(c+1)}] \quad . \quad (5)$$

Theorem: The Rule minimizes

$$E\{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{.i} - \mu_{.i})^2\}$$

subject to

$$E\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{.i}\} = P \quad .$$

Proof: The proof is by induction. We first show that a randomized rule applied to any additional case will result in a larger mean square error than our rule. We then assume that our rule is better than randomizing any r cases and show that it is also better than randomizing any (r+1) cases. The details are contained in Huang [9].

The average mean square error of classification is given by

$$E\{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{.i} - \mu_{.i})^2\} = \sum_{j=1}^{c-1} F_{(j)}P_{(j)} + F_{(c)}[A + P_{(c)} - 2AP_{(c)}] + \sum_{j=c+1}^{8} F_{(j)}[1-P_{(j)}] \quad (6)$$

Estimation of Parameters

In the practical situation the parameters of interest must be estimated from the sample data. We denote the eight sample proportions by

$$\hat{P}_{000}, \hat{P}_{100}, \hat{P}_{010}, \hat{P}_{110}, \hat{P}_{001}, \hat{P}_{101}, \hat{P}_{011}, \hat{P}_{111} . \quad (7)$$

There are five independent parameters, say p_{1u}, p_{2u}, P, α, R, the remaining parameters being defined by identities and the unbiasedness restrictions. Define $\theta' = (p_{1u}, p_{2u}, P, \alpha, R)$, $f(\theta) = (P_{000}, P_{100}, P_{010}, P_{110}, P_{001}, P_{101}, P_{011})'$ and $Y = (\hat{P}_{000}, \hat{P}_{100}, \hat{P}_{010}, \hat{P}_{110}, \hat{P}_{001}, \hat{P}_{101}, \hat{P}_{011})'$ then we may express the observed proportions as

$$Y = f(\theta) + e , \quad (8)$$

where $E\{e\} = 0$. The covariance matrix of e is that of the multinomial with parameters $f(\theta)$. The Gauss-Newton method of estimation (see Fuller [4] and Hartley [8]) may then be used to solve this non-linear regression problem.

Example

The Statistical Laboratory of Iowa State University in cooperation with the Statistical Reporting Service of the U.S. Department of Agriculture conducted a survey of 262 Iowa farm operators in September and October of 1970. In both of these interviews the respondents were asked to name the most important product of their farm operation. A good deal of information on the farm operation was also collected. We consider the variables

Y_{mi} = 1 if the i^{th} farm operator reports hogs the most important product on the m^{th} interview, m = 1,2
= 0 otherwise

X_{3i} = 1 if the number of breeding hogs is equal to or greater than 30 for the i^{th} farm operator,
= 0 otherwise.

The analysis is summarized in Table 1. The estimated parameters obtained by the Gauss-Newton procedure are $(\tilde{p}_{1u}, \tilde{p}_{2u}, \tilde{P}, \tilde{\alpha}, \tilde{R})$ = (0.9153, 0.9159, 0.3890, 0.6716, 0.3585). The estimated standard errors for these estimators are (0.0357, 0.0356, 0.0281, 0.0498, 0.0295).

In this particular example the two methods of obtaining the information are two identical questions asked at different times. Therefore we would expect the values of p_{1u} and p_{2u} to be about the same. The estimates for these parameters are approximately equal.

Using the estimates the conditional probability that the true value is 1 is estimated for each cell and is given on the fourth line of the table. On the basis of these estimated conditional probabilities we assign the estimated individual true value, $\tilde{\mu}_{.i}$, as follows:

$\tilde{\mu}_{.i}$ = 1 if individual i belongs to case (1,1,1), (1,1,0), (0,1,1) or (1,0,1),

= 1 for a random sample of 4 of the 13 individuals who belong to the case (0,1,0),

= 0 otherwise.

The estimated mean square error of this classification, $\hat{MSE}(\tilde{\mu}_{.i})$, is 0.0439. If we use only the first question the estimated classification error is 0.0659 and if we use the second question alone the estimated classification error is 0.0654. Thus the use of two questions and the auxiliary information has reduced the estimated classification error by about one third.

Acknowledgement

This research was partially supported by Joint Statistical Agreement J.S.A. 73-2 with the Bureau of the Census, U. S. Department of Commerce.

REFERENCES

[1] Bryson, M. R., "Errors of classification in a binomial population," Journal of the American Statistical Association 60 (1965), 217-224.

[2] Cochran, W. G., "Problems arising in the analysis of a series of experiments," Journal of the Royal Statistical Society Supplement 4 (1937), 102-118.

[3] Cochran, W. G. and Carroll, S. P., "A sampling investigation of the efficiency of weighting inversely as the estimated variance," Biometrics 9 (1953), 447-459.

[4] Fuller, W. A., "Gauss-Newton estimation with a preliminary estimate," Unpublished class notes, Department of Statistics, Iowa State University, 1972.

[5] Giesbrecht, F. G., "Classification errors and measures of association in contingency tables," Proceedings of the Social Statistics Section of the American Statistical Association 1967, 271-276.

[6] Hansen, M. H., Hurwitz, W. N. and Bershad, M. A., "Measurement errors in censuses and surveys," Bulletin de Institute International de Statistique 38, No. 2 (1961), 359-374.

[7] Hansen, M. H., Hurwitz, W. N., Marks, E. S. and Mauldin, W. P., "Response errors in surveys," Journal of the American Statistical Association 46 (1951), 147-190.

[8] Hartley, H. O., "The Gauss-Newton method for the fitting of non-linear regression functions by least squares,: Technometrics 3 (1961), 269-280.

[9] Huang, H. T., "Combining multiple responses in sample surveys," Unpublished Ph.D. dissertation, Iowa State University, 1972.

[10] Koch, G. G., "The effect of non-sampling errors on measures of association in 2 x 2 contingency tables," Journal of the American Statistical Association 64 (1969), 852-863.

[11] Meier, P., "Variance of a weighting mean," Biometrics 9 (1953), 59-73.

[12] Mote, V. L. and Anderson, R. L., "An investigation of the effects of misclassification on the properties of χ^2-tests in the analysis of categorical data," Biometrika 52 (1965), 95-109.

Table 1. Most Important Product (Example)

(Y_{1i}, Y_{2i}, X_{3i})	(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,1)	Total
Obs. Frequency	116	8	13	32	29	3	8	53	262
Obs. Proportion	0.4428	0.0305	0.0496	0.1221	0.1107	0.0115	0.0305	0.2023	1.0000
Est. Model Prob.	0.4428	0.0366	0.0365	0.1257	0.1071	0.0245	0.0246	0.2022	1.0000
Est. Cond. Prob. that $\mu = 1$	0.0024	0.3120	0.3154	0.9887	0.0160	0.7552	0.7581	0.9983	-
$\tilde{\mu}_{.i}$	0	0	0.3288*	1	0	1	1	1	-

*Randomization probability