CLASSIFICATION ERRORS AND MEASURES OF ASSOCIATION IN CONTINGENCY TABLES*

F. G. Giesbrecht, Research Triangle Institute

It has become a well recognized fact that statisticians must be cognizant of both sampling and non-sampling errors in the analysis and interpretation of data obtained from sample surveys. For the purpose of this study, the term response error is rather loosely defined to include all effects which result in an incorrect classification in the final tabulations. These can be due to such diverse sources as deliberate falsification by the respondent or incorrect recording by the interviewer.

The purpose of the model described in this paper is to provide a means for investigating the effect of response errors on selected measures of association in contingency tables and to aid in the design of special surveys for the purpose of estimating these errors. The model for 2 x 2 contingency tables contains a total of 13 parameters, including three basic probabilities and ten response error parameters. The response error parameters are defined as conditional probabilities. The two characteristics will be referred to as A and B with the respective complements being $\bar{A}$, (not A) and $\bar{B}$ (not B). Consequently an individual is identified as belonging to both A and B, i.e. AB, A and not B i.e. $A\bar{B}$, not A but B i.e. $\bar{A}B$ or finally neither A nor B and denoted by $\bar{A}\bar{B}$. These four classes are disjoint and exhaustive.

Let P_B be the probability that a randomly selected individual belongs to class B. Let $P_{A|B}$ be the conditional probability that a randomly selected individual from class B also belongs to class A. Similarly $P_{A|\bar{B}}$ is the conditional probability that a randomly selected individual who is not in class B, is in class A. It follows that the probability that a randomly selected individual will belong to A and not to B is equal to $P_{A|\bar{B}}(1-P_B)$. Probabilities for the other three possibilities have equivalent definitions. The special case in which $P_{A|B} = P_{A|\bar{B}}$ is the one in which there is no association between the two factors.

Response Error Parameters

The three basic parameters defined in the previous section would be sufficient if there were only sampling errors. However, the actual classification (abbreviated as ac) will at times differ from the true classification (abbreviated as tc). Now define the response error parameters:

β_1 = Pr (ac is B | tc is B).

β_0 = Pr (ac is $\bar{B}$ | tc is $\bar{B}$).

These two probabilities do not depend on the A classification. A slightly more flexible model can be obtained by introducing four probabilities for errors in the B classification and allowing a dependence on the actual A classification.

α_{11} = Pr (ac is AB | tc is AB and ac is B).

α_{01} = Pr (ac is $\bar{A}B$ | tc is $\bar{A}B$ and ac is B).

α_{10} = Pr (ac is AB | tc is $A\bar{B}$ and ac is B).

α_{00} = Pr (ac is $\bar{A}B$ | tc is $\bar{A}\bar{B}$ and ac is B).

γ_{11} = Pr (ac is $A\bar{B}$ | tc is AB and ac is $\bar{B}$).

γ_{01} = Pr (ac is $\bar{A}\bar{B}$ | tc is $\bar{A}B$ and ac is $\bar{B}$).

γ_{10} = Pr (ac is $A\bar{B}$ | tc is $A\bar{B}$ and ac is $\bar{B}$).

γ_{00} = Pr (ac is $\bar{A}\bar{B}$ | tc is $\bar{A}\bar{B}$ and ac is $\bar{B}$).

Since a randomly selected individual can belong to any one of four classes and be assigned to any one of four classes, the parameters define the likelihood of the 16 distinct possibilities. The probability that a randomly selected individual will be assigned to class AB is

$$\beta_1\alpha_{11}P_{A|B}P_B + (1-\beta_0)\alpha_{10}P_{A|\bar{B}}(1-P_B) + \beta_1(1-\alpha_{01})(1-P_{A|B})P_B + (1-\beta_0)(1-\alpha_{00})(1-P_{A|\bar{B}})(1-P_B) .$$

Similarly the probability a randomly selected individual will be assigned to class $A\bar{B}$ is

$$(1-\beta_1)\gamma_{11}P_{A|B}P_B + \beta_0\gamma_{10}P_{A|\bar{B}}(1-P_B) + (1-\beta_1)(1-\gamma_{01})(1-P_{A|B})P_B + \beta_0(1-\gamma_{00})(1-P_{A|\bar{B}})(1-P_B) .$$

The probability of being assigned to class $\bar{A}B$ is

$$\beta_1(1-\alpha_{11})P_{A|B}P_B + (1-\beta_0)(1-\alpha_{10})P_{A|\bar{B}}(1-P_B) + \beta_1\alpha_{01}(1-P_{A|B})P_B + (1-\beta_0)\alpha_{00}(1-P_{A|\bar{B}})(1-P_B) .$$

Finally, the probability of being assigned to class $\bar{A}\bar{B}$ is

* Research supported by Bureau of the Census, Contract No. Cco-9191.

$$(1-\beta_1)(1-\gamma_{11})P_{A|B}P_B + \beta_0(1-\gamma_{10})P_{A|\bar{B}}(1-P_B)$$

$$+ (1-\beta_1)\gamma_{01}(1-P_{A|B})P_B$$

$$+ \beta_0\gamma_{00}(1-P_{A|B})(1-P_B) .$$

Effect of Response Errors on the Chi-square Statistic

A study of the non-centrality parameter of the χ^2 test statistic when there is no association, that is $P_{A|B} = P_{A|\bar{B}}$ and β_0 and β_1 are the only response error parameters not equal to one, verifies the statement by Bross that errors of this type do not disturb the validity of the χ^2 test. However, these errors do decrease the non-centrality parameter when $P_{A|B} \neq P_{A|\bar{B}}$. This is the phenomenon of loss of power of the χ^2 test in the presence of response errors. However, response errors do not always have these effects. For example, if $\alpha_{11} = .9$, $P_B = .5$, $P_{A|B} = .9$, $P_{A|\bar{B}} = .9$ and all other parameters equal one, then the non-centrality parameter is not zero even though there is no true association and hence an invalid test. For this case the non-centrality parameter turns out to be .016. However if in the above case, $P_{A|B} = .7$, then the non-centrality parameter is .101. This can be compared to .062 when there are no response errors. The implication is an increase in power. Other cases can be examined in a similar manner. Similar techniques can be used to study the effects of combinations of response errors on other measures of association.

Estimation of the Response Error Parameters

Since it is reasonably simple to examine the effects of the response error parameters, the interesting problem is to estimate these parameters. Assume that a second interviewer is assigned the task of reinterviewing a random sample of those individuals already surveyed. Based on the responses to characteristics A and B in both interviews, each individual will be assigned to one of 16 classes. These classes are combinations of the four possible assignments as a result of the first interview and the four following the second interview. If the two interviews are assumed to be independent and that the same response error parameters apply to both interviews, then the expected values of the fraction in each of the 16 categories are as given in Table I. The task now is to determine which functions of the parameters are estimable. An examination of Table I shows immediately that the expected number assigned to the class A and $\bar{B}$ by the first interviewer and to class AB by the second interviewer is equal to the expected number assigned to class AB by the first interviewer and to class A and $\bar{B}$ by the second interviewer. There are five other pairs with matching expected values. Also the sum of all 16 frequencies is equal to unity, implying a maximum of 9 degrees of freedom for purposes of estimation. However, there are 13 parameters in the model. It is obvious that not all parameters are estimable. The problem now is to find which parameters or functions of the parameters are estimable. The method for locating these is an application of the definition of an estimable function, that is, a function is estimable if there exists a function which estimates it. These functions are located by equating the observed relative frequencies to the expected values and then solving the resulting equations for meaningful functions of the parameters. The estimators obtained in this manner may not be optimal in any sense. It is simply a verification that the function can be estimated. Once it has been verified that a set of functions of the parameters is estimable, one can use any of the standard methods, such as maximum likelihood or minimum chi-square to obtain estimates with desirable properties. If one of these methods is chosen, one will need to use one of the iterative, numerical techniques to arrive at the final solutions.

An example of the method for verifying that certain parameters are estimable is as follows:

The expected value of the sum of the four classes for which both interviewers have recorded that the individual belongs to class B is

$$\beta_1^2 P_B + (1-\beta_0)^2 (1-P_B) .$$

Similarly the expected value of the sum of those recorded as $\bar{B}$ by both interviewers is

$$(1-\beta_1)^2 P_B + \beta_0^2 (1-P_B) .$$

This leads to two equations in three unknowns. If one of the three is known and P_B does not have an extreme value then one can solve for the remaining two. Hence two of the three are estimable.

Further applications of this technique lead to other estimable functions. Unfortunately the set of estimable functions obtained in this manner is not unique, but rather is a function of the assumptions one is willing to make. For example, the above derivation illustrates that if one knows β_0 then β_1 and P_B are estimable. Alternatively if one assumes $\beta_0 = \beta_1 = \beta$ then β and P_B are estimable. The appropriate choice for any given situation depends on the supporting information available from other sources.

It can be shown that if it is assumed that if β_1, α_{01}, α_{00}, γ_{01} and γ_{00} are known and β_1 is not equal to an extreme value, then P_B, $P_{A|B}$, $P_{A|\bar{B}}$, β_0, α_{11}, α_{10}, γ_{11} and γ_{10} are estimable. Alternatively, if it is assumed that $\beta_0 = 1$ and that $\alpha_{11} = \alpha_{01}$, $\alpha_{10} = \alpha_{00}$, $\gamma_{11} = \gamma_{10}$, and $\gamma_{01} = \gamma_{00}$, then β_1, P_B, $P_{A|B}$, $P_{A|\bar{B}}$, α_{11}, α_{10}, γ_{11} and γ_{01} are estimable.

ACKNOWLEDGEMENTS

The author acknowledges the benefit derived from several informal discussions at the Bureau of the Census. Participants included Max A. Bershad, William N. Hurwitz, Tom Jabine, Leon Pritzker and Dr. Benjamin Tepping. The author also acknowledges the assistance and encouragement by Dr. D. G. Horvitz of the Research Triangle Institute.

REFERENCES

[1] Assakul, Kwanchai and C. H. Proctor, "Testing Hypotheses with Categorical Data Subject to Misclassification," N. C. Inst. of Statistics, Mimeograph Series No. 448, Raleigh, N. C., (1965).

[2] Bross, Irwin, "Misclassification in 2 x 2 Tables." Biometrics 10, (1954), 478-486.

[3] Mote, V. L., "An Investigation of the Effects of Misclassification on the X^2 tests in the Analysis of Categorical Data," N. C. Inst. of Statistics, Mimeograph Series No. 182, Raleigh, N. C., (1957).

Table I Expected Values of the Frequencies of Various Types of Classification in Two Independent Interviews

	Classification: First Interview	Classification: Second Interview	Fraction	Expected Value
1	AB	AB	$f(A_1B_1A_2B_2)$	$\beta_1^2\alpha_{11}^2 P_{A\|B}P_B + (1-\beta_0)^2\alpha_{10}^2 P_{A\|\bar{B}}(1-P_B) + \beta_1^2(1-\alpha_{01})^2(1-P_{A\|B})P_B$ $+ (1-\beta_0)^2(1-\alpha_{00})^2(1-P_{A\|\bar{B}})(1-P_B)$
2	AB	$A\bar{B}$	$f(A_1B_1A_2\bar{B}_2)$	$\beta_1(1-\beta_1)\alpha_{11}\gamma_{11}P_{A\|B}P_B + \beta_0(1-\beta_0)\alpha_{10}\gamma_{10}P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1(1-\beta_1)(1-\alpha_{01})(1-\gamma_{01})(1-P_{A\|B})P_B + \beta_0(1-\beta_0)(1-\alpha_{00})(1-\gamma_{00})(1-P_{A\|\bar{B}})(1-P_B)$
3	$A\bar{B}$	AB	$f(A_1\bar{B}_1A_2B_2)$	$\beta_1(1-\beta_1)\alpha_{11}\gamma_{11}P_{A\|B}P_B + \beta_0(1-\beta_0)\alpha_{10}\gamma_{10}P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1(1-\beta_1)(1-\alpha_{01})(1-\gamma_{01})(1-P_{A\|B})P_B + \beta_0(1-\beta_0)(1-\alpha_{00})(1-\gamma_{00})(1-P_{A\|\bar{B}})(1-P_B)$
4	$A\bar{B}$	$A\bar{B}$	$f(A_1\bar{B}_1A_2\bar{B}_2)$	$(1-\beta_1)^2\gamma_{11}^2P_{A\|B}P_B + \beta_0^2\gamma_{10}^2P_{A\|\bar{B}}(1-P_B) + (1-\beta_1)^2(1-\gamma_{01})^2(1-P_{A\|B})P_B$ $+ \beta_0^2(1-\gamma_{00})^2(1-P_{A\|\bar{B}})(1-P_B)$
5	AB	$\bar{A}B$	$f(A_1B_1\bar{A}_2B_2)$	$\beta_1^2\alpha_{11}(1-\alpha_{11}) P_{A\|B}P_B + (1-\beta_0)^2\alpha_{10}(1-\alpha_{10}) P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1^2\alpha_{01}(1-\alpha_{01})(1-P_{A\|B})P_B + (1-\beta_0)^2\alpha_{00}(1-\alpha_{00})(1-P_{A\|\bar{B}})(1-P_B)$
6	AB	$\bar{A}\bar{B}$	$f(A_1B_1\bar{A}_2\bar{B}_2)$	$\beta_1(1-\beta_1)\alpha_{11}(1-\gamma_{11}) P_{A\|B}P_B + \beta_0(1-\beta_0)\alpha_{10}(1-\gamma_{10}) P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1(1-\beta_1)(1-\alpha_{01})\gamma_{01}(1-P_{A\|B})P_B + \beta_0(1-\beta_0)(1-\alpha_{00})\gamma_{00}(1-P_{A\|\bar{B}})(1-P_B)$

Table I Continued

7	$A\bar{B}$	$\bar{A}B$	$f(A_1\bar{B}_1\bar{A}_2B_2)$	$\beta_1(1-\beta_1)(1-\alpha_{11})\gamma_{11}\ P_{A\|B}P_B + \beta_0(1-\beta_0)(1-\alpha_{10})\gamma_{10}P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1(1-\beta_1)\alpha_{01}(1-\gamma_{01})(1-P_{A\|B})P_B + \beta_0(1-\beta_0)\alpha_{00}(1-\gamma_{00})(1-P_{A\|\bar{B}})(1-P_B)$
8	$A\bar{B}$	$\bar{A}\bar{B}$	$f(A_1\bar{B}_1\bar{A}_2\bar{B}_2)$	$(1-\beta_1)^2\gamma_{11}(1-\gamma_{11})P_{A\|B}P_B + \beta_0^2\gamma_{10}(1-\gamma_{10})P_{A\|\bar{B}}(1-P_B)$ $+ (1-\beta_1)^2\gamma_{01}(1-\gamma_{01})(1-P_{A\|B})P_B + \beta_0^2\ \gamma_{00}(1-\gamma_{00})(1-P_{A\|\bar{B}})(1-P_B)$
9	$\bar{A}B$	AB	$f(\bar{A}_1B_1A_2B_2)$	$\beta_1^2\alpha_{11}(1-\alpha_{11})P_{A\|B}P_B + (1-\beta_0)^2\alpha_{10}(1-\alpha_{10})P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1^2\alpha_{01}(1-\alpha_{01})(1-P_{A\|B})P_B + (1-\beta_0)^2\alpha_{00}(1-\alpha_{00})(1-P_{A\|\bar{B}})(1-P_B)$
10	$\bar{A}\bar{B}$	AB	$f(\bar{A}_1\bar{B}_1A_2B_2)$	$\beta_1(1-\beta_1)\alpha_{11}(1-\gamma_{11})P_{A\|B}P_B + \beta_0(1-\beta_0)\alpha_{10}(1-\gamma_{10})P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1(1-\beta_1)(1-\alpha_{01})\gamma_{01}(1-P_{A\|B})P_B + \beta_0(1-\beta_0)(1-\alpha_{00})\gamma_{00}(1-P_{A\|\bar{B}})(1-P_B)$
11	$\bar{A}B$	$A\bar{B}$	$f(\bar{A}_1B_1A_2\bar{B}_2)$	$\beta_1(1-\beta_1)(1-\alpha_{11})\gamma_{11}P_{A\|B}P_B + \beta_0(1-\beta_0)(1-\alpha_{10})\gamma_{10}P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1(1-\beta_1)\alpha_{01}(1-\gamma_{01})(1-P_{A\|B})P_B + \beta_0(1-\beta_0)\alpha_{00}(1-\gamma_{00})(1-P_{A\|\bar{B}})(1-P_B)$
12	$\bar{A}\bar{B}$	$A\bar{B}$	$f(\bar{A}_1\bar{B}_1A_2\bar{B}_2)$	$(1-\beta_1)^2\gamma_{11}(1-\gamma_{11})\ P_{A\|B}P_B + \beta_0^2\gamma_{10}(1-\gamma_{10})\ P_{A\|\bar{B}}(1-P_B)$ $+ (1-\beta_1)^2\gamma_{01}(1-\gamma_{01})(1-P_{A\|B})P_B + \beta_0^2\gamma_{00}(1-\gamma_{00})(1-P_{A\|\bar{B}})(1-P_B)$
13	$\bar{A}B$	$\bar{A}B$	$f(\bar{A}_1B_1\bar{A}_2B_2)$	$\beta_1^2(1-\alpha_{11})^2\ P_{A\|B}P_B + (1-\beta_0)^2(1-\alpha_{10})^2P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1^2\alpha_{01}^2(1-P_{A\|B})P_B + (1-\beta_0)^2\alpha_{00}^2(1-P_{A\|\bar{B}})(1-P_B)$

Table I Continued

14	$\bar{A}\bar{B}$	$\bar{A}B$	$f(\bar{A}_1\bar{B}_1\bar{A}_2B_2)$	$\beta_1(1-\beta_1)(1-\alpha_{11})(1-\gamma_{11})\ P_{A\|B}P_B + \beta_0(1-\beta_0)(1-\alpha_{10})P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1(1-\beta_1)\alpha_{01}\gamma_{01}(1-P_{A\|B})P_B + \beta_0(1-\beta_0)\alpha_{00}\gamma_{00}(1-P_{A\|\bar{B}})(1-P_B)$
15	$\bar{A}B$	$\bar{A}\bar{B}$	$f(\bar{A}_1B_1\bar{A}_2\bar{B}_2)$	$\beta_1(1-\beta_1)(1-\alpha_{11})(1-\gamma_{11})P_{A\|B}P_B + \beta_0(1-\beta_0)(1-\alpha_{10})(1-\gamma_{10})P_{A\|\bar{B}}(1-P_B)$ $+ \beta_1(1-\beta_1)\alpha_{01}\gamma_{01}(1-P_{A\|B})P_B + \beta_0(1-\beta_0)\alpha_{00}\gamma_{00}(1-P_{A\|\bar{B}})(1-P_B)$
16	$\bar{A}\bar{B}$	$\bar{A}\bar{B}$	$f(\bar{A}_1\bar{B}_1\bar{A}_2\bar{B}_2)$	$(1-\beta_1)^2(1-\gamma_{11})^2P_{A\|B}P_B + \beta_0^2(1-\gamma_{10})^2P_{A\|\bar{B}}(1-P_B) + (1-\beta_1)^2\gamma_{01}^2(1-P_{A\|B})P_B$ $+ \beta_0^2\gamma_{00}^2(1-P_{A\|\bar{B}})(1-P_B)$