A GENERAL RESPONSE AND MEASUREMENT ERROR MODEL AND ITS APPLICATION TO THE ANALYSIS OF 2 x 2 CONTINGENCY TABLES[1/]

A. Lawrence Gould, Research Triangle Institute

1. Introduction

Nonsampling errors in survey data have been the subject of considerable interest and research, as the extensive bibliography in Cochran [1] indicates. Errors of this sort may be classed as nonresponse errors, where for one reason or another observed measurements on some of the sample units are not available, and what are often called "measurement" errors; this paper is concerned with these latter errors.

Treatments of "measurement" errors tend to take one of two approaches: a model is postulated for their probability structure and the consequences with regard to the distributions of various statistics are worked out; or, starting with reinterview data, various statistics are devised to indicate the effects of the "measurement" errors, with little explicit specification of a probability structure for the errors. The research discussed is an attempt to tie these approaches together by developing a general model for response and measurement errors, and showing how the elements of this model enter into the distribution of statistics developed through the second approach just mentioned.

2. The General Model

Suppose there is a population of N discrete individuals, and that with each individual there is associated a vector, ξ, of characteristics. Usually, for an individual, the elements of ξ are regarded as being fixed; however, there are circumstances in which it is more appropriate to regard them as random variables. If the elements of ξ are regarded as fixed, the measurement errors have the interpretation that on repeated attempts to observe the value of ξ, different realized values would occur. The elements of ξ may be regarded appropriately as random when, for instance, they correspond to states of mind of an individual with regard to various political or social issues. When a survey is conducted at some time (or trial) t, and a sampled individual is approached for interview or measurement, ξ assumes a particular value ξ_t. The distribution of potential values of ξ may be different for different individuals, although they will be assumed to be in the same family of distributions, so that the individual distributions will differ only in the values of the elements of a parameter vector, θ. The distribution (probability or density function) of ξ over all trials t, for the i-th population member will be denoted by $f(\xi;\theta_i)$. In turn, θ may be regarded as having a distribution over different individuals of the population, with probability or density function $g(\theta;\Theta)$, Θ being a set of parameters characteristic of the population. The act of drawing a sample of individuals from the population is therefore equivalent to drawing a sample of θ's, and also equivalent to drawing a sample of ξ distributions.

When a survey is conducted under conditions γ at trial t, the i-th member of the population (if in the sample) will yield a measured response η_{it} when information regarding ξ is elicited. If the measuring or interviewing process contains no errors, then $\eta_{it} = \xi_{it}$, the value of ξ at trial t for the i-th individual. This is, however, only the ideal situation; in practice, η_{it} may be regarded as a random variable with a distribution depending upon ξ_{it} and the conditions under which the survey is conducted. The probability or density function of η_{it} (the evoked response of individual i at trial t may be denoted by

$$h_\gamma(\gamma;\ \xi_{it},\ \phi_i(\gamma)) \quad ,$$

where $\phi_i(\gamma)$ denotes a set of parameters specific to the i-th individual and survey conditions γ. As do the θ's, the ϕ's have a distribution over the population, with probability or density function $k_\gamma(\phi, \Phi(\gamma))$. The notion of a "trial" t may be generalized to include a set of "trials" $\{t_1,\ldots,\ t_m\}$, in which case ξ_{it} and η_{it} are vectors of possibly correlated elements:

$$\xi_{it} = (\xi_{it_1},\ldots,\ \xi_{it_m})'$$

$$\eta_{it} = (\eta_{it_1},\ldots,\ \eta_{it_m})'.$$

Thus, repeated surveys may readily be dealt with in the framework of the model.

The distribution $f(\xi;\ \theta_i)$ may be thought of as the response error distribution for an individual because it describes the distribution of potential responses (ξ's) the individual might present at an interview. The distribution $h_\gamma(\eta;\ \xi_{it},\ \phi_i(\gamma))$ may be regarded as the (conditional) measurement error distribution which depends upon the survey conditions, the individual, and the response he presents at the interview; it describes the distribution of potential measurements of the response. For this reason, ξ will be called the "response", and η, the "measurement".

As a physical example to illustrate the distinction between these two conceptual errors, suppose one wishes to measure the length of a metal bar on two occasions, and that the unit of length is taken to be the meter bar formerly used as a length standard as it existed at a given instant of time. Then, at either occasion, the metal bar of interest has, in terms of this length standard, a definite length. The objective of the measurement procedure is to determine what the lengths on the two oc-

1/ Research supported by the U. S. Bureau of the Census, Contract No. Cco-9319.

casions are. Because of environmental differences between the two occasions of measurement, the "true" length of the bar may not be the same on both occasions; this variation is the response error. On measuring the bar's length on either occasion, the inaccuracy of the measuring instruments may lead to an observed length different from the "true" length at that occasion; this variation is the "measurement" error.

As well as the distribution of measurements on an individual conditional upon a particular response ξ_t, one may envisage a distribution for the measurements averaged over all possible responses under a given set of survey conditions. Such a distribution is

$$h^*_\gamma(\eta;\ \theta_i,\ \phi_i(\gamma)) = \begin{cases} \int_\xi h_\gamma(\eta;\ \xi,\ \phi_i(\gamma))\ f(\xi;\ \theta_i)\ d\xi \\ \\ \sum_\xi h_\gamma(\eta;\ \xi,\ \phi_i(\gamma))\ f(\xi;\ \theta_i) \end{cases},$$

depending upon whether ξ is a continuous or discrete random variable. Intuitively, h^* may be regarded as follows: given a particular individual, if one could obtain a measurement η at various trials under the same survey conditions and the act of measuring did not affect the process generating the measurements, then the observed distribution of measurements would be described by h^*_γ.

The distributions f, h, and h* are marginal distributions. In reality it may happen that the responses or measurements of different individuals are correlated; this is essentially the point at which the two approaches to analyzing non-sampling errors of the "measurement" type diverge. Generally speaking, population structure approaches assume the between-individual correlations to be either absent or to have a particular form; the sample statistic-oriented approach makes no assumption about the absence of correlations between individuals.

The remainder of this paper is concerned with an application of the model to the analysis of 2 x 2 contingency tables. In this situation, an individual is in, or is assigned to, one of four classes according to his possession or non-possession of either of two attributes. For notational convenience, an individual's response and measurement may be regarded as 4-element vectors; the j-th element of a vector being unity if and only if the individual is in, or is assigned to, the j-th of the four possible cells of the contingency table, and the remaining elements of the vector being zero.

3. An Example of the Population Structure Approach

From the population structure point of view, the roles of the response and measurement error distributions are so defined as to yield, for one or more trials, a presumed distribution for the observed cell frequencies. In application, then, the problem becomes that of estimating the parameters appearing in the functional form of the presumed distribution.

As a simple example of a population structure model for the 2 x 2 contingency table situation, assume that observed classifications of different individuals are independent, as are the classifications of any individual on separate trials. Then

$$f(\xi;\ \theta_i) = \xi'\theta_i\ ,$$

$$h(\eta;\ \xi_i,\ \phi_i) = \eta'\phi_i\xi_i\ ,$$

$$h^*(\eta;\ \theta_i,\ \phi_i) = \eta'\phi_i\theta_i$$

(ϕ is a matrix; all the other entities are vectors).

Suppose θ_i and ϕ_i are independently distributed, and that $\theta^* = E(\theta)$ and $\phi^* = E(\phi)$; then the averages of the frequency distributions f and h* over all individuals in the population are

$$f'(\xi;\ \theta^*) = \xi'\theta^*$$

and

$$h^{*\prime}(\eta;\ \theta,^*\ \phi^*) = \eta'\phi^*\theta^*\ .$$

On the average basis, h*' may be taken as the distribution of the various cells in the table, and the distribution of observed frequencies is a multinomial distribution with the four possible values of h*' as the parameters. The objective is, given a set of observed classifications, to estimate the values of the elements of ϕ^* and θ^*. If the sample units are classified on two occasions, and the responses of the individuals are assumed not to change between the two trials, then the average joint distribution of the cells is

$$h^{*\prime}(\eta_1,\ \eta_2;\theta^*,\ \phi^*) = \eta_1'\ \phi^* \begin{bmatrix} \theta_1^* & & 0 & \\ & \theta_2^* & & \\ & & \theta_3^* & \\ 0 & & & \theta_4^* \end{bmatrix} \phi^{*\prime}\eta_2\ .$$

The models developed by Giesbrecht [3] and by Koch [6] lead to distributions which are special cases of (1) in that additional assumptions regarding the elements of ϕ^* are imposed. The population structure model developed in the example used could be generalized by altering the independence assumptions.

4. An Example of the Sample-Oriented Approach

The sample-oriented approach is by and large concerned with estimating the population proportions falling into the various categories, and with investigating the roles of various sources of error and various intercorrelations on the precision of the estimators of these proportions. Suppose that the elements of Θ represent the true proportions of the population falling into the various cells of the table, where Θ is defined by

$$\Theta = E\{\xi\} = \sum_\theta \sum_\xi \xi\ f(\xi;\ \theta)\ g(\theta;\ \Theta)\ .$$

In the presence of measurement errors, an unbiased estimate of Θ may not be obtainable. The expected value of a single observation on individual i of the population is

$$E\{\eta_i\} = \Theta + \sum_{\theta}\sum_{\xi}\left(\sum_{\eta} \eta\, h_{\gamma}(\eta_i;\xi,\phi_i(\gamma)) - \xi\right) \cdot f(\xi;\theta)\, g(\theta;\Theta) \quad ;$$

in general, the value of the bias term is not known. Given a simple random sample of n units from the population, the usual estimator of Θ is $\bar{Y}$, the sample mean. Other sampling schemes could be employed as well; the computations with simple random sampling are probably the simplest. The covariance matrix of the elements of $\bar{Y}$ may be shown to have the following form, assuming the θ's and ϕ's to be independently distributed in the population:

$$V\{\bar{Y}\} = \frac{n+(N-1)a}{nN} \cdot \frac{1}{N}\sum_{i=1}^{N}\{\phi^*\theta_i\,\theta_i' - \mathrm{diag}\,[(\phi^*\theta_i)_1,\ldots,(\phi^*\theta_i)_4]\} + \frac{a}{nN}\sum_{i=1}^{N}\phi^*(\theta_i-\theta^*)(\theta_i-\theta^*)'\phi^{*\prime} + \frac{n-a}{n}\left(1-\frac{1}{N}\right)\cdot\frac{1}{N(N-1)} \cdot \sum_{i\neq i'}\sum E\{(\eta_i - \phi^*\theta_i)(\eta_{i'} - \phi^*\theta_{i'})'\} \qquad (2)$$

where $\theta^* = E\{\phi\}$, $\phi^* = E\{\phi\}$, and the distribution of η_i is h^*_{γ}; if the sampling is without replacement, $a = (N-n)/(N-1)$; if the sampling is with replacement, $a = 1$; in both cases, the sampling is simple random. The quantities appearing in (2) correspond to those defined in the example in the previous section. The first term of (2) arises from the combined effects of response and measurement errors; the second term reflects the sampling error, and the third term expresses the degree of correlation between observed classifications of individuals -- it is proportional to what is usually called the "within-trial correlation between different individuals," -- or at least would be if the quantities of interest were scalars rather than vectors.

In a practical situation, the objective is often to estimate the sampling and response-measurement contributions. For the present case, it may be shown that, if $N \ll n$ and the third term of (2) is zero (which is not, by the way, to say that the measurement of an individual is uncorrelated with that of an other individual), then the usual estimator of the covariance matrix of the elements of the sample mean is an approximately unbiased estimator of $V\{\bar{Y}\}$. Further, if a repetition of the survey is conducted under conditions identical to those of the original survey, and if measurements on different trials are assumed uncorrelated, then the quantity g,

$$g = \sum_{j=1}^{n}(Y_{jt} - Y_{jt'})(Y_{jt} - Y_{jt'})' \quad ,$$

where Y_{jt} is the observed measurement on the j-th sample unit, with expectation

$$E\{g\} = \frac{n}{N}\sum_{i=1}^{N}\left\{\mathrm{diag}[(\phi^*\theta_i)_1,\ldots,(\phi^*\theta_i)_4] - \phi^*\theta_i\theta_i'\phi^{*\prime}\right\} \quad ,$$

provides the basis for an unbiased estimator of the response-measurement component of (2), namely $g/2n^2$. Thus, the moments of the quantities considered in the sample-oriented approach may be related to the distribution appearing in the general model.

The example discussed in this section is based upon Koch's [7] extension to the 2 x 2 contingency table case of the work of Hansen, Hurwitz, and Pritzker [5] which, in turn, is based on the model of Hansen, Hurwitz, and Bershad [4]. Felligi [2] discusses the application of the Hansen, Hurwitz, and Bershad model to interpenetrating sample and re-interview surveys, considering in detail the various intercorrelations which might arise.

5. Remarks

One purpose of the proposed model is to provide a conceptual way of considering non-sampling errors of the "measurement" type which allows the population-structure and sample-oriented approaches to considering these errors to be related to each other, with an explicit indication of the respects in which the two approaches differ. For the 2 x 2 contingency table case, this relating of the two approaches is fairly straightforward under some simplifying assumptions, as the preceding discussion indicates.

The examples presented by no means exhaust the possibilities for developing the general model, and a number of areas of extension may readily be visualized, for instance: considering the applicability of the model to a number of other measurement error models, including those discussed by Cochran and by Mandel [8]; extension of the general model to include the possibility of sampling plans more general than simple random sampling, or surveys in which only a subset of the sample is reinterviewed, or to include the effect of non-response, etc.

6. References

[1] Cochran, W. G. (1968). Errors of measurement in statistics. Technometrics 40, 637-666.

[2] Felligi, I. P. (1964). Response variance and its estimation. Jour. Am. Stat. Assoc. 59, 1016-1041.

[3] Giesbrecht, F. G. (1967). Evaluation of the effects of classification errors on measures of association in contingency tables. RTI Report SU-215, prepared for Bureau of Census, Contract Cco-9191.

[4] Hansen, M. H., Hurwitz, W. N., and Bershad, M. A. (1961). Measurement errors in cen-

suses and surveys. Bull. Int. Stat. Inst. 38, No. 2, 359-374.

[5] Hansen, M. H., Hurwitz, W. N., and Pritzker, L. (1964). The estimation and interpretation of gross differences and the simple response variance. Contributions to Statistics Presented to Professor P. C. Mahalanobis, Pergamon Press, Oxford.

[6] Koch, G. G. (1968). A simple model for misclassification errors in 2 x 2 contingency tables. RTI report SU-363, prepared for Bureau of Census, Contract Cco-9260.

[7] Koch, G. G. (1968). The effect of nonsampling errors on measures of association in 2 x 2 contingency tables. RTI Report SU-363, prepared for Bureau of Census, Contract Cco-9260.

[8] Mandel, J. (1959). The measuring process. Technometrics 1, 251-267.