ON INFERENCE FROM GENERAL CATEGORICAL DATA WITH MISCLASSIFICATION ERRORS BASED ON DOUBLE SAMPLING SCHEMES

by

Yosef Hochberg

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1066

April 1976

On Inference From General Categorical Data With Misclassification Errors Based on Double Sampling Schemes

Yosef Hochberg

Summary:

In order to resolve the difficulties involved in inference from a sample of categorical data obtained by using a fallible classifying mechanism (usually inexpensive), we consider the utilization of a sub-sample subjected to a simultaneous cross-classification of its elements by the fallible mechanism and by some true (usually expensive) classifying mechanism. The setup is general; i.e., the discussion can be applied to any multidimensional cross-classified data obtained by unrestricted random sampling. Two methodologies are presented: (i) Maximum likelihood approach, (ii) Least squares approach. Both methodologies are illustrated using real data.

Introduction.

Much has been written on the effects of misclassification errors on studies of association in 2×2 contingency tables. We refer the reader to Fleiss (1973, Ch. 11) for a review of that subject. However, very little has been accomplished in developing techniques for resolving the problems of misclassification errors even for the relatively simple setup of studying association in a 2×2 contingency table.

In Koch (1969) the misclassification errors are assumed to be generated according to a random response error model. As such, the

methodology is based on repeated classifications of the experimental elements. Such a methodology cannot always be satisfactory because of obvious practical difficulties and because in many problems, misclassification errors are fixed biases rather than random response errors.

A model for fixed bias misclassification errors has been introduced by Bross (1954) for a 2 × 2 table where two error parameters are assumed. However, in a 2 × 2 table one may theoretically have as many as 12 different parameters for fixed bias misclassification errors, e.g. the probability of classifying an element in the first row and the second column when the element actually belongs in the second row and the first column, etc. Usually studies of the potential effects of fixed bias misclassification errors were limited to cases where most error parameters were assumed to be zero. Even for these special cases, no methodologies have been presented for statistical inference. In Hochberg and Reinfurt (1975), the effects of incorporating six error parameters on three measures of association in 2 × 2 tables are discussed.

In this paper, we extend the use of a double sampling scheme (originally introduced by Tenenbein (1970), (1971), (1972)) for inference (i.e. modeling, testing hypotheses etc.) from general multi-dimensional contingency tables with 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. The reader should not adhere to the mechanical connotation of the term "device." The experimental setup referred to is very often met in reality in problems where the

distinction between a true or a false classification device simply relates to making or not making an extra effort to obtain good data. Such experimental situations are frequently met by researchers in various domains of science. For example, Diamond and Lilienfeld (1962) discuss an experimental situation in public health research where the true classification device is the physician's examination whereas the fallible classifier is a questionnaire completed by the patient. Another real experimental situation will be discussed in the sequel to exemplify the procedures of the following sections.

In real problems it is often the case that the true classification device uses different nominal scales than those used by the fallible device. The experimenter's knowledge of the degree of correspondence between the levels of two such scales may vary from none to complete. For the preceding example, a nominal scale for a patient's response to a questionnaire may have four levels A, B, C, and D while the physician's report may use some standard scale with levels 1, 2, 3, 4, 5, and 6. The correspondence between the patient's scale and the physician's scale may be clear (i.e. A $\leftrightarrow$ (1 or 2), B $\leftrightarrow$ (3 or 4), C $\leftrightarrow$ 5, D $\leftrightarrow$ 6) or, as more often is the case, it may be quite unclear. Note that we refer to correspondence between these scales as implied by the apriori definitions of the scales and their levels. Even in cases where such a relation is completely known, fixed bias errors of misclassification may very well prevail. The procedures to be discussed here have double motivation in such experimental situations. First, they can be used to resolve the problems of misclassification errors. Secondly, even when misclassification errors do not exist, the procedures enable one to carry out an efficient study expressing results in terms of the finer scale utilized in a

relatively small sub-sample.

In Section 2 we describe the setup and the notation to be used. In Section 3, the maximum likelihood approach is discussed and in Section 4, we examine the least squares approach. Section 5 contains a practical technique for using the maximum likelihood approach. In Section 6 we give some examples.

The discussion in Sections 2 through 5 will pertain to a setup where all variables are subjected to misclassification errors when the fallible device is used. In practice one might come across situations when only a subset of the variables is subjected to errors. Two cases are of interest. One case is when the magnitudes of errors within combinations of levels of the correctly reported set of variables are possibly different and the other case is when these errors can be assumed to be the same across the corresponding levels. The first case is readily obtained along the line of our following discussion. We do not treat this case in detail but we do, however, use the examples section to demonstrate it. The second case is not readily obvious. We hope to discuss that and related problems in a following work.

2. The Setup and Notation.

Two independent samples are drawn from the target population. Each is an unrestricted simple random sample. If the actual frame for the population is finite we adhere to the concept of a 'super' population. See e.g. Hartley and Sielken (1975). The first sample of n_1 elements is only classified by the fallible device. Let $\underset{\sim}{i}' = (i_1', \ldots, i_d')$ index a specific combination of levels of the d variables under study. The second sample of n_2 elements is simultaneously classified by both the false and the true devices. Here again, we use $\underset{\sim}{i}'$ to index the fallible

cell. To index the true classification we use $\underset{\sim}{i} = (i_1,\ldots,i_d)$. Also let $i_m = 1,\ldots,I_m$ and $i'_m = 1,\ldots,I'_m$, $m = 1,\ldots,d$ with $I_1 \times I_2 \times \ldots \times I_d = k$ and $I'_1 \times I'_2 \times \ldots \times I'_d = k'$.

Next we introduce notation for the frequencies and parameters in the two samples. To simplify matters we use the same letters to indicate similar conceptual quantities in both samples. The distinction, however, will be easy since the second sample will always have two indices corresponding to true and false classifications respectively. Thus, $n(\underset{\sim}{i}')$ denotes the frequency in the $\underset{\sim}{i}'$-th cell as obtained in the first sample by the false classifier. Similarly, $n(\underset{\sim}{i},\underset{\sim}{i}')$ denotes the frequency in the second sample classified in the $\underset{\sim}{i}$-th cell by the true classifier and in the $\underset{\sim}{i}'$-th cell on the false categorical scale. Likewise, let $\gamma(\underset{\sim}{i}')$ and $\gamma(\underset{\sim}{i},\underset{\sim}{i}')$ denote the corresponding population proportions. We introduce $\gamma(\underset{\sim}{i}|\underset{\sim}{i}') = \gamma(\underset{\sim}{i},\underset{\sim}{i}')/\gamma(\underset{\sim}{i}')$, which is the fraction of times an element actually belongs to cell $\underset{\sim}{i}$ when reported to be in cell $\underset{\sim}{i}'$ by the fallible classifier. The convention of replacing an index by a period to indicate that summation has been taken over that index will be used throughout, e.g. $n(\underset{\sim}{i},\cdot) = \Sigma_{\underset{\sim}{i}'} n(\underset{\sim}{i},\underset{\sim}{i}')$.

The intermediate parameters of interest are clearly the $\gamma(\underset{\sim}{i},\cdot)$ for which we use the special notation $\gamma(\underset{\sim}{i},\cdot) \equiv \pi(\underset{\sim}{i})$.

Throughout this work we will use the convention of putting an under bar to indicate a vector. An indexed vector will be used only for the $\gamma(\underset{\sim}{i}|\underset{\sim}{i}')$ where $\underset{\sim}{\gamma}(\underset{\sim}{i}') = (\gamma(\underset{\sim}{i}|\underset{\sim}{i}'),$ for all $\underset{\sim}{i})$.

3. Inference Based on Maximum Likelihood Estimates (MLE) of the $\pi(\underset{\sim}{i})$.

Given the data, the likelihood function of the $\gamma(\underset{\sim}{i}')$ and the $\gamma(\underset{\sim}{i}|\underset{\sim}{i}')$ is given by (see next page)

$$F = a \prod_{\underset{\sim}{i}'} [\gamma(\underset{\sim}{i}')]^{n(\underset{\sim}{i}')+n(\cdot,\underset{\sim}{i}')} \prod_{\underset{\sim}{i}'} b(\underset{\sim}{i}') \prod_{\underset{\sim}{i}} [\gamma(\underset{\sim}{i}|\underset{\sim}{i}')]^{n(\underset{\sim}{i},\underset{\sim}{i}')},$$

where -a- is a constant depending on the n_i, $i = 1,2$, the $n(\underset{\sim}{i}')$, and the $n(\cdot,\underset{\sim}{i}')$; the $b(\underset{\sim}{i}')$ are constants depending on the $n(\cdot,\underset{\sim}{i}')$ and the $n(\underset{\sim}{i},\underset{\sim}{i}')$. It is now easily verified that the MLE's are given by:

$$\hat{\gamma}(\underset{\sim}{i}|\underset{\sim}{i}') = \frac{n(\underset{\sim}{i},\underset{\sim}{i}')}{n(\cdot,\underset{\sim}{i}')}$$

$$\hat{\gamma}(\underset{\sim}{i}') = \frac{n(\underset{\sim}{i}') + n(\cdot,\underset{\sim}{i}')}{n_1 + n_2} \,. \tag{3.1}$$

Since the $\pi(\underset{\sim}{i})$ and the $\gamma(\underset{\sim}{i}|\underset{\sim}{i}')$ are in 1:1 relation with the set of $\gamma(\underset{\sim}{i})$ and $\gamma(\underset{\sim}{i}|\underset{\sim}{i}')$, the MLE's of the $\pi(\underset{\sim}{i})$ are given by:

$$\hat{\pi}(\underset{\sim}{i}) = \sum_{\underset{\sim}{i}'} \hat{\gamma}(\underset{\sim}{i}')\gamma(\underset{\sim}{i}|\underset{\sim}{i}'), \ \forall \underset{\sim}{i} \tag{3.2}$$

Next we consider the asymptotic variance matrix of $\hat{\underset{\sim}{\pi}}$ which we denote by $V(\hat{\underset{\sim}{\pi}})$. Note that, asymptotically, the set of the $\hat{\gamma}(\underset{\sim}{i}')$ is independent of the set of $\hat{\gamma}(\underset{\sim}{i}|\underset{\sim}{i}')$. A similar statement applies to any two distinct vectors $\underset{\sim}{\gamma}(\underset{\sim}{i}')$, $\underset{\sim}{\gamma}(\underset{\sim}{j}')$, $\underset{\sim}{i}' \neq \underset{\sim}{j}'$. This is clear from the block diagonal information matrix which is easily obtained from F. Linearizing the $\hat{\pi}(\underset{\sim}{i})$ by a Taylor approximation around the $\gamma(\underset{\sim}{i}')$ and $\gamma(\underset{\sim}{i}|\underset{\sim}{i}')$, we get (for large sample sizes)

$$\hat{\pi}(\underset{\sim}{i}) \simeq \sum_{\underset{\sim}{i}'} \gamma(\underset{\sim}{i}|\underset{\sim}{i}')\hat{\gamma}(\underset{\sim}{i}') + \sum_{\underset{\sim}{i}'} \gamma(\underset{\sim}{i}')\hat{\gamma}(\underset{\sim}{i}|\underset{\sim}{i}').$$

On letting

$$V(\hat{\underset{\sim}{\gamma}}) = \frac{1}{n_1 + n_2}\left[D(\underset{\sim}{\gamma}) - \underset{\sim}{\gamma}\,\underset{\sim}{\gamma}'\right] \equiv ((v_{m,n})), m,n = 1,\ldots,k'$$

$$V\left[\hat{\underset{\sim}{\gamma}}(\underset{\sim}{i}')\right] = \frac{1}{n_2\gamma(\underset{\sim}{i}')}\left[D(\underset{\sim}{\gamma}(\underset{\sim}{i}')) - \underset{\sim}{\gamma}(\underset{\sim}{i}')\underset{\sim}{\gamma}'(\underset{\sim}{i}')\right] \equiv V_{\underset{\sim}{i}'}$$

where $D(\cdot)$ is a diagonal matrix with the vector $(\cdot)$ displayed on the main diagonal, we have asymptotically

$$V(\underset{\sim}{\hat{\pi}}) = \sum_{m=1}^{k'}\sum_{n=1}^{k} v_{m,n}\underset{\sim}{\gamma}(\underset{\sim}{i}'_m)\underset{\sim}{\gamma}(\underset{\sim}{i}'_n) + \sum_{\underset{\sim}{i}'}\gamma^2(\underset{\sim}{i}')V_{\underset{\sim}{i}'} . \tag{3.3}$$

When consistent estimators from (3.1) are substituted for the $\underset{\sim}{\gamma}(\underset{\sim}{i}')$ and the $\gamma(\underset{\sim}{i}')$ in (3.3), one obtains a consistent estimator $\hat{V}(\underset{\sim}{\hat{\pi}})$ for the dispersion matrix of the vector $\underset{\sim}{\hat{\pi}}$.

A Maximum Likelihood Test of fit is simple. (We take the hypothesis of fit here to be the hypothesis: $\gamma(\cdot,\underset{\sim}{i}') = \gamma(\underset{\sim}{i}')$ for all $\underset{\sim}{i}'$). The unrestricted MLE's are given by:

$$\hat{\gamma}(\underset{\sim}{i}') = \frac{n(\underset{\sim}{i}')}{n_1}$$

$$\hat{\hat{\gamma}}(\underset{\sim}{i},\underset{\sim}{i}') = \frac{n(\underset{\sim}{i},\underset{\sim}{i}')}{n_2}$$

Under the hypothesis, the MLE's of the $\gamma(\underset{\sim}{i},\underset{\sim}{i}')$ are $\hat{\gamma}(\underset{\sim}{i},\underset{\sim}{i}') = \hat{\gamma}(\underset{\sim}{i}|\underset{\sim}{i}')\hat{\gamma}(\underset{\sim}{i}')$. On denoting the Maximum Likelihood Ratio (MLR) statistic by L, we have

$$-2\log L = -2\left\{\sum_{\underset{\sim}{i}'} n(\underset{\sim}{i}')\log\left[\frac{\hat{\hat{\gamma}}(\underset{\sim}{i}')}{\hat{\gamma}(\underset{\sim}{i}')}\right] + \sum_{\underset{\sim}{i}}\sum_{\underset{\sim}{i}'} n(\underset{\sim}{i},\underset{\sim}{i}')\log\left[\frac{\hat{\hat{\gamma}}(\underset{\sim}{i},\underset{\sim}{i}')}{\hat{\gamma}(\underset{\sim}{i},\underset{\sim}{i}')}\right]\right\}.$$

Under the restrictive hypothesis this is asymptotically distributed as a central Chi-square variate with k-1 d.f.

Often, having established the fit, the experimenter will be interested in further inference on $\underset{\sim}{\pi}$ based on the efficient estimator $\underset{\sim}{\hat{\pi}}$. In most practical problems it is unfeasible to obtain simple MLE's of $\underset{\sim}{\pi}$ under further functional restrictions on the $\pi(\underset{\sim}{i})$ (given the model fits). One can verify this by trying to obtain the MLR test for independence in a 2 × 2 table. Even for this simple problem the MLE

cannot be obtained explicitly and one must call upon numerical techniques. In general, the usual log-linear hypotheses on $\underset{\sim}{\pi}$ (hypotheses given by $C\underset{\sim}{\pi} = \underset{\sim}{0}$ or $C\log(\underset{\sim}{\pi}) = \underset{\sim}{0}$, where C is some contrast matrix, i.e. $C'\underset{\sim}{1} = \underset{\sim}{0}$) will impose complicated functional relationships among the $\gamma(\underset{\sim}{i}|\underset{\sim}{i}')$ and the $\gamma(\underset{\sim}{i}')$. The MLE's will have to be obtained by some numerical computer subroutines.

The practical approach is to utilize the estimator $\hat{\underset{\sim}{\pi}}$ and the consistent estimator of its variance matrix, $\hat{V}(\hat{\underset{\sim}{\pi}})$, as initial input to the asymptotically valid least squares procedures presented in Grizzle et al (1969) and Forthofer and Koch (1973). This is discussed at more length in Section 5 where a convenient technique is given for implementing the Maximum Likelihood approach in first stage and proceeding with weighted least squares approach with one computer program.

4. Inference Based on Least Squares Estimators (LSE) of the $\pi(\underset{\sim}{i})$.

Before starting our discussion (which will resemble to some extent that in Koch et al, (1972)), the notation must be altered slightly. We let $p(\underset{\sim}{i}') = n(\underset{\sim}{i}')/n_1$ and $p(\underset{\sim}{i},\underset{\sim}{i}') = n(\underset{\sim}{i},\underset{\sim}{i}')/n_2$. Let $\underset{\sim}{p}_1$ be the vector whose elements are all $p(\underset{\sim}{i}')$. Similarily, let $\underset{\sim}{p}_2$ be the vector of length k'k obtained by stretching out all the $\underset{\sim}{p}(\underset{\sim}{i}')$ in order. Finally, let $\underset{\sim}{\gamma}_i = E(\underset{\sim}{p}_i)$, i = 1,2 and denote $\underset{\sim}{p} = (\underset{\sim}{p}_1', \underset{\sim}{p}_2')'$.

The dispersion matrix of $\underset{\sim}{p}$ is a block diagonal matrix $V(\underset{\sim}{p})$ with $V(\underset{\sim}{p}_i)$ on the diagonal, where

$$V(\underset{\sim}{p}_i) = \frac{1}{n_i}\left[D(\underset{\sim}{\gamma}_i) - \underset{\sim}{\gamma}_i\underset{\sim}{\gamma}_i'\right], \quad i = 1,2.$$

Next write $\underset{\sim}{F} = A\underset{\sim}{p}$,

$$A = \begin{bmatrix} A_1 : (k'-1) \times k' & \vdots & O \\ \cdots\cdots\cdots & \vdots & \cdots\cdots\cdots \\ O & \vdots & A_2 : (k'k-1) \times k'k \end{bmatrix}$$

where A_1 is obtained from an identity matrix of dimension k' by deleting the last row and A_2 is similarly obtained from an identity matrix of dimension k'k. We can now write a model

$$E(\underset{\sim}{F}) = X\underset{\sim}{\theta},\ \underset{\sim}{\theta} : (k'k-1) \times 1. \tag{4.1}$$

where $X = (X_1', X_2')'$ with X_2 being an identity matrix of order k'k-1 and X_1 of dimensions $(k'-1) \times (k'k-1)$ has the form $X_1 = [I \otimes \underset{\sim}{1}' \mid 0]$ where I is the identity matrix of order k'-1, $\underset{\sim}{1}' = (1,\ldots,1)$ and $\otimes$ denotes Kroncker's Delta multiplication.

The variance matrix of $\underset{\sim}{F}$ is consistently estimated by $V(\underset{\sim}{\hat{F}}) = A\hat{V}(\underset{\sim}{p})A'$ where $\hat{V}(\underset{\sim}{p})$ is obtained when substituting the unrestricted MLE's of the $\underset{\sim}{V}_i$ in the expression for $V(\underset{\sim}{p})$. Thus, in large samples one may use weighted least squares to estimate the vector $\underset{\sim}{\theta}$. The asymptotic LSE of $\underset{\sim}{\theta}$ (which is BAN if (4.1) holds) is given by

$$\underset{\sim}{\hat{\theta}} = (X'\hat{V}^{-1}(\underset{\sim}{F})X)^{-1}X'\hat{V}^{-1}(\underset{\sim}{F})\underset{\sim}{F}$$

and the consistent estimator of its dispersion matrix $V(\underset{\sim}{\hat{\theta}})$ is given by

$$\hat{V}(\underset{\sim}{\hat{\theta}}) = (X'\hat{V}^{-1}(\underset{\sim}{F})X)^{-1}.$$

A test for goodness of fit is based on

$$X^2 = \underset{\sim}{F}'\hat{V}^{-1}(\underset{\sim}{F})\underset{\sim}{F} - \underset{\sim}{\hat{\theta}}'X'\hat{V}^{-1}(\underset{\sim}{F})X\underset{\sim}{\hat{\theta}}$$

which, under the hypothesis that the model fits, follows an asymptotic Chi-square distribution with k'-1 d.f.

If the model adequately describes the data, tests of hypotheses with respect to the parameters comprising $\underset{\sim}{\theta}$ can be undertaken. Note

that the elements of $\underset{\sim}{\theta}$ are the k'k-1 upper-left elements among the k'k parameters $\gamma(\underset{\sim}{i},\underset{\sim}{i}')$. The last element is obtained from the relation $\Sigma_{\underset{\sim}{i}}\Sigma_{\underset{\sim}{i}'}\gamma(\underset{\sim}{i},\underset{\sim}{i}') = 1$. From $\hat{\underset{\sim}{\theta}}$ and its estimated variance matrix one can easily obtain the LSE of $\underset{\sim}{\pi},\tilde{\underset{\sim}{\pi}}$ say, and its estimated variance matrix $\hat{V}(\tilde{\underset{\sim}{\pi}})$. Inference based on this approach is exemplified in Section 6.

5. Employing the Maximum Likelihood Approach.

Here we use notations from both Sections 3 and 4. The MLE's of the $\pi(\underset{\sim}{i})$ and their asymptotic variance matrix were given in Section 3. The overall procedure of first obtaining MLE's and then using asymptotic Least Squares theory appears somewhat inconvenient especially when considering the available computer programs. Here we discuss a simple technique to implement the methodology of Section 3 which can be employed using a single computer program.

This approach is based on the fact that the MLE's of the $\pi(\underset{\sim}{i})$ can be expressed as compound exponential-logarithmic-linear functions (see Forthofer and Koch, 1973) of the elements of $\underset{\sim}{p}$.

Specifically, we can write $\hat{\underset{\sim}{\pi}}$ (the MLE of $\underset{\sim}{\pi}$, see Section 3) as

$$\hat{\underset{\sim}{\pi}} = Q\left\{\exp[K\log(A\underset{\sim}{p})]\right\}$$

where

$$\underset{(2+k)k' \times (k+1)k'}{A} = \left[\begin{array}{c:c} a_1 I:k' \times k' & a_2 I \otimes \underset{\sim}{1}':k' \times k'k \\ \hdashline 0:k' \times k' & I \otimes \underset{\sim}{1}:k' \times k'k \\ \hdashline 0:k'k \times k' & I:k'k \times k'k \end{array}\right]$$

$a_i = n_i/(n_1 + n_2)$, $i = 1,2$, and

$$\underset{k'k \times k'(k+2)}{K} = [I \otimes \underset{\sim}{1}, -I \otimes \underset{\sim}{1}, I{:}k'k \times k'k]$$

where the unspecified identity matrix I has dimension k' and $\underset{\sim}{1}$ is of length k. Finally, the matrix Q is given by

$$\underset{k \times k'k}{Q} = I \otimes 1'.$$

Thus, on letting $\underset{\sim}{y} = A\underset{\sim}{p}$ and $z = \exp[K\log(\underset{\sim}{y})]$ we can conveniently write the asymptotic variance matrix of $\hat{\underset{\sim}{\pi}}$ as

$$V(\hat{\underset{\sim}{\pi}}) = QD(\underset{\sim}{z})KD^{-1}(\underset{\sim}{y})AV(\underset{\sim}{p})A'D^{-1}(\underset{\sim}{y})K'D(\underset{\sim}{z})Q',$$

where the generic notation $D(\cdot)$ is as defined in Section 2.

As noted earlier, the vector $\hat{\underset{\sim}{\pi}}$ and the estimated variance matrix $\hat{V}(\hat{\underset{\sim}{\pi}})$ (which is obtained by substituting $\hat{V}(\underset{\sim}{p})$ for $V(\underset{\sim}{p})$) are subsequently used as initial inputs for further modeling based on weighted (asymptotic) least squares procedures as in Grizzle et al (1969). Thus, one may obtain functions of the $\hat{\underset{\sim}{\pi}}$ which are of interest for further modeling via a repeated chain of linear, log or exponential transformations and then express a linear model for the resulting functions. The model can be tested for fit and if it fits, linear hypotheses on its estimable parameters can be tested. All this can be done by a single computer run using the new program GENCAT given in Landis and Stanish (1975). This is exemplified in our next section.

6. Examples.

Here we present several examples which came up recently in highway safety research. The problem is that of evaluating the effectiveness of safety belts (lap and shoulder) in reducing injuries in automobile accidents. One of the difficulties in assessing the

effectiveness of belts is the fact that most information on injuries and belt usage by occupants is obtained from the police reports on the corresponding accidents. Workers in highway safety research have recently become aware that there are systematic misclassification errors in the police reports with respect to both occupant injury and belt usage. The possible effects of such biases on three measures of belt effectiveness (measures of association between belt usage and injury)were evaluated in Hochberg and Reinfurt (1975). As is apparent from that study, even small misclassification errors can lead to an enormous inflation or deflation of the true values depending on the magnitudes of the specific true values and the error parameters.

Another difficulty in obtaining "true" measures for safety belt effectiveness in reducing injury has to do with the fact that wearers and non-wearers of belts do not have similar distributions over the levels of certain relevant factors, e.g., type of car (make, size, etc.), type of driver (age, sex, physical condition, etc.), environment, and accident type (rollover, head-on collision, etc.).

In order to resolve these difficulties we adopted the methodology given above. A double sampling scheme was outlined in which the "true" classification mechanism of individuals was based on intensive inquiries for each individual case. These intensive inquiries included detailed reports from hospitals on injuries, belt usage, etc., for those individuals who were taken to a hospital. Individuals who did not go to a hospital were contacted by telephone to obtain parallel information. The fallible classification "device" here is the police-reported classification.

First we look at the overall effectiveness of using only lap belts across all factors of interest. In this setup we only look at the variables belt-usage with two levels--"yes" and "no" and injury again with two levels--"yes" and "no". For our original data (first sample of size n_1) we have the total police-reported accidents in the state of North Carolina during the year 1974. The corresponding breakdown is given in Table 1.

Table 1. (Belt usage)×(injury) in 1974 North Carolina accidents by police report

		Usage		
		No	Yes	Total
Injury	No	123625	16325	139950
	Yes	25780	2622	28402
	Total	149405	18947	168352

The injury rates are: No belt - 17.255%

Lap belt - 13.839%

The difference in risk is thus 3.416% which gives an estimated relative effectiveness for lap belts of 19.80%.

Our supplementary sample had 2142 valid observations when cross classified by police and non-police sources on both belt usage and injury. The data for this breakdown is given in Table 2, where "P" and "NP" indicate "police and "non-police" respectively.

In this case we assume that errors of misclassification exist in the police reported injury and the police reported belt usage and that these errors for each of the two variables may be different within different levels of the other variable. This assumption can be confirmed from our data.

Only for this example we provide the specifications for using the procedures discussed above.

Table 2. (Usage-P)×(injury-P)×(usage-NP)×(injury-NP)

N-P / P	No Injury No Belt	No Injury Lap Belt	Injury No Belt	Injury Lap Belt	Total
No Injury No Belt	1194	143	227	28	1592
No Injury Lap Belt	14	133	5	19	171
Injury No Belt	28	5	306	11	350
Injury Lap Belt	2	3	3	21	29
Total	1238	284	541	79	2142

When using GENCAT for either our Maximum-Likelihood (ML) or Least-Squares (LS) procedures the data for this run is entered as a vector of frequencies for two populations (original and supplementary samples) each with sixteen responses as follows:

(Frequency data vector input to GENCAT)' = (123625 0 0 0

0 16325 0 0 0 0 25780 0 0 0 0 2622 1194 143 227 28 14 133

5 19 28 5 306 11 2 3 3 21)'

ML Estimates. Denote the vector of proportions formed from the data input by $\underset{\sim}{p}$ (note the slight difference between this $\underset{\sim}{p}$ and the $\underset{\sim}{p}$ in former sections). First we transform $\underset{\sim}{p}$ linearly to $A_1\underset{\sim}{p}$ where

$$A_1 \underset{24 \times 32}{=} \begin{bmatrix} a_1 I \otimes \underset{\sim}{1}' & a_2 I \otimes \underset{\sim}{1}' \\ O & I \otimes \underset{\sim}{1}' \\ \cdots\cdots & \\ I{:}16 \times 16 & \end{bmatrix}$$

, where the unspecified I and $\underset{\sim}{1}$ are of dimension 4 here, and $a_i = n_i/(n_1 + n_2)$, i = 1,2.

Next $A_1\underset{\sim}{p}$ is transformed to $\log(A_1\underset{\sim}{p})$ and on to $A_2\log(A_1\underset{\sim}{p})$ where

$$\underset{16 \times 24}{A_2} = [I \otimes \underset{\sim}{1} \;:\; I \otimes -\underset{\sim}{1} \;:\; I{:}16 \times 16]$$

We now exponentiate the resulting vector and multiply it from the left by the matrix

$$\underset{4 \times 16}{A_3} = \begin{bmatrix} 0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0 \\ 1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0 \\ 0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1 \\ 0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&0 \end{bmatrix}$$

Next we take the log transformation and then a linear operator from the left by

$$\underset{2 \times 4}{A_4} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

Next we exponentiate the vector and the resulting 2×1 vector contains the ML estimates of incidence of injury when not using the belt and when using the lap belt respectively. Here:

	No Belt	Lap Belt
Injury Percentage:	29.825%	21.510%

The estimated covariance matrix is given by:

$$\begin{bmatrix} .7414 & -.0878 \\ -.0878 & 3.8231 \end{bmatrix}$$

The differential risk is thus estimated by 8.315% with an estimated standard error = 2.177%.

LS Estimates. The vector $\underset{\sim}{p}$ is first linearly transformed from left by

$$\underset{18 \times 32}{A_1^*} = \begin{bmatrix} I \otimes \underset{\sim}{1}' \\ \cdots\cdots \\ I{:}15 \times 15 \end{bmatrix}$$

, where the unspecified I and $\underset{\sim}{1}$ are of dimension 4.

To the resulting vector, $\underset{\sim}{F}$, a linear model is fitted as in Grizzle et al, (1969),

$$E(\underset{\sim}{F}) = X\,\underset{\sim}{\theta}\ , \text{ where}$$

$$X_{15 \times 18} = \begin{bmatrix} 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0 \\ - - - - - - - - - - - - - - - - \\ \multicolumn{18}{c}{I:15 \times 15} \end{bmatrix}$$

From the LSE of $\underset{\sim}{\theta}$ and its estimated covariance matrix one forms (cannot be done within one computer run when using GENCAT) the LS estimates of the "true" proportions and their estimated covariance matrix. These are then used as initial inputs to a second stage run where first the "true" proportions are transformed by the linear operator

$$A_2^* = \begin{bmatrix} 0\ 0\ 1\ 0 \\ 1\ 0\ 1\ 0 \\ 0\ 0\ 0\ 1 \\ 0\ 1\ 0\ 1 \end{bmatrix}$$

then by log and then linear by $A_3^* = A_4$. After exponentiating the resulting vector one obtains the LS estimates. In this case we get

	No Belt	Lap Belt
Injury Percentage:	29.826%	21.512%

and the estimated covariance matrix is

$$\begin{bmatrix} .7613 & -0.0918 \\ -0.0918 & 3.3210 \end{bmatrix}$$

Note the similarity in results from the MS and LS procedures.

The next example involves a similar problem, but on a larger scale. In the previous example, only two variables were considered (belt usage and injury). In addition to these variables, this example will consider two other variables or factors, namely, vehicle damage severity (low, high) and driver sex (male, female).

The framework assumed here is that of no misclassification errors in reporting levels of damage severity and driver sex. We do not assume that the misclassification errors in police reports of belt usage and injury are the same for all levels of damage severity by sex of driver.

Tables 3 and 4 present the data on which later calculations are based. Table 3 gives frequency of police reported belt usage by injury for each damage/sex category, while Table 4 compares non-police and police reports of belt usage by injury for these same categories. In order to use the GENCAT program, this data is entered as in the former example; i.e., all frequencies of Table 3 are systematically entered first, followed by those of Table 4.

Table 3. Frequencies of (belt usage) x (injury) within (damage) x (sex) in 1974 North Carolina accidents based on police reports.

	Damage Severity (low)				Damage Severity (high)			
	Male		Female		Male		Female	
	None	Lap	None	Lap	None	Lap	None	Lap
No Injury	22536	3006	11199	1262	17476	2155	6964	728
Injury	1687	199	1422	117	6746	583	3707	297

Table 4. Frequencies of (usage - P) x (injury - P) x (usage - NP) x (injury - NP) within (damage) x (sex).

			Non-Police															
			Damage Severity (low)								Damage Severity (high)							
			Male				Female				Male				Female			
			No Injury		Injury		No Injury		Injury		No Injury		Injury		No Injury		Injury	
			None	Lap	None	Lap	None	Lap	None	Lap	None	Lap	None	Lap	None	Lap	None	Lap
Police	No Injury	None	299	20	59	9	102	7	53	4	407	62	45	7	206	18	37	5
		Lap	4	33	1	6	2	6	1	3	6	47	1	6	1	17	0	1
	Injury	None	11	2	118	5	5	1	79	1	5	1	32	4	4	0	29	0
		Lap	1	2	0	9	0	0	1	6	0	1	1	2	0	0	1	0

Also, as in the former example, transformations are made within each of the four damage by sex combinations, to obtain the MLE's of the injury probabilities for both belted and unbelted drivers within each of these four cells. Next, the ratio of (injury rate for unbelted)/(injury rate for belted) is computed for each (driver sex) x (damage severity) category. These "true" measures of relative risk, based on police data that have been adjusted by supplementary non-police information, are given in Table 5. Their (asymptotic) covariance matrix follows.

Table 5. Estimated relative risks.

Damage Severity (low)		Damage Severity (high)	
Male	Female	Male	Female
1.109	1.078	1.154	1.918

$$\begin{bmatrix} .0259 & .0225 \times 10^{-9} & -.1040 \times 10^{-9} & -.5750 \times 10^{-9} \\ & .0627 & .0302 \times 10^{-9} & .1667 \times 10^{-9} \\ & \text{(symmetric)} & .0473 & -.7713 \times 10^{-9} \\ & & & .5725 \end{bmatrix}$$

In order to describe the variation between the four (driver sex) x (damage severity) cells, a saturated model was fitted and the interaction effect between sex and damage severity on seat belt effectiveness tested. The resulting Chi-square with one degree of freedom was 0.89, which gave a p-value of 0.35.

Next, both main effects were simultaneously tested by using the contrast matrix

$$C = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$$

The resulting Chi-square in this case was 1.14 (two d.f.), with a p-value of 0.57. These "non-significant" results are most likely due to the relatively small amount of data from the supplementary sample, compared with the original police report data.

In an upcoming report, similar data will be analyzed where the inference will be based on an initial model which summarizes the variability in the misclassification errors in a subset of the variables across the remaining variables.

REFERENCES

Bross, I. (1954). Misclassification in 2x2 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.

Fleiss, J.L. (1973). Statistical Methods for Rates and Proportions. New York: Wiley, Inc.

Forthofer, R.M. and Koch, G.G. (1973). An anlysis for compounded functions of categorical data. Biometrics, 29, 143-157.

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

Hartley, H.O. and Sielken, R.L., Jr. (1975). A "super-population viewpoint" for finite population sampling. Biometrics, 31, 411-422.

Hochberg, Y. and Reinfurt, D.W. (1975). An investigation of safety belt usage and effectiveness. NHTSA Project (DOT HS-4-00897) Interim Report.

Koch, G.G. (1969). The effect of nonsampling errors on measure of association in 2x2 contingency tables. Journal of the American Statistical Association, 65, 851-864.

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.

Landis, R.J. and Stanish, W.M. (1975). GENCAT: A Computer Program for the Generalized Chi-Squares Analysis of Categorical Data Using Weighted Least Squares to Compute Wald Statistics. Department of Biostatistics, University of North Carolina,Chapel Hill.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. J. Amer. Statis. Assoc., 65, 1350-1361.

Tenenbein, A. (1971). A double sampling scheme for estimating from binomial data with misclassifications; sample size determination. Biometrics, 27, 935-944.

Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multimonial data with application to sampling inspection. Technometrics, 14, 187-202.