Biometrical Journal **44** (2002) 5, 603–618

Generalizing Double and Triple Sampling for Repeated Surveys and Partial Verification

JOHN M. HOENIG, R. CHOUDARY HANUMARA, and DENNIS M. HEISEY

Virginia Institute of Marine Science
College of William and Mary

Summary

Population composition is often estimated by double sampling in which the value of a covariate is noted on each of a large number of randomly selected units and the value of the covariate and the exact class to which the unit belongs is noted for a smaller sample. The cross-classified sample can be used to estimate the classification rates and these, in turn, can be used in conjunction with the estimated distribution of the covariate to obtain an improved estimate of the population composition over that obtained by direct observation of the identity of the individuals in a small sample. There are two approaches to this problem characterized by the way in which the classification rates are defined. The simplest approach uses estimates of the probability $P(i \mid j)$ that the unit is actually in class i given that the covariate is in class j. The more complicated approach uses estimates of the probability $P(j \mid i)$ that the covariate falls in class j given that the unit is actually in class i. The latter approach involves estimating more parameters than the former but avoids the necessity for the two samples to be drawn from the same population. We show the two approaches can be combined when there are multiple surveys. For example, one might conduct a disease survey for several years; in each year the accurate and/or error-prone techniques may be applied to samples. The sensitivities and specificities of the error-prone test are assumed constant across surveys. Generalizations allow for more than one error-prone classifier and partial verification (estimation of misclassification rates by application of the accurate technique to fixed subsamples from each error-prone category). The general approach is illustrated by considering a repeated survey for malaria.

Key words: Disease surveys; Error-prone tests; Misclassification probabilities; Contingency tables; Age-length keys; Stratified random sampling; Fallible classifiers.

1. Introduction

It is often logistically or economically impractical to measure the value of a primary variable of interest on a large enough sample. This has led to so-called double sampling, where a more readily observed covariate is observed on a large sample. Because the covariate does not perfectly reflect the value of the primary variable, it is necessary to obtain a small sample on which both the primary variable and the covariate are observed so that the association between the variables can be characterized.

 0323-3847/02/0507-0603 $ 17.50+.50/0

This problem occurs in numerous guises, as illustrated by the following three examples. 1) In a disease survey, one may use an inexpensive but error-prone test to examine a large number of animals. On a much smaller sample, one might use both the error-prone test and an expensive but exact test to determine the misclassification rates of the error-prone test. This information can then be used to correct the results obtained from that part of the study in which only the error-prone test was used. 2) In fisheries research, it is important to estimate the age composition of the catch. A common procedure is to measure the lengths of a large number of fish as this information is easy to obtain. On a smaller sample, the otoliths (ear stones) are collected and the ages of the fish are determined by counting annual growth rings. From knowledge of the length composition of the catch and the age composition within each length class it is possible to estimate the age composition of the catch. 3) In the social sciences, a great deal of information may be collected inexpensively by having people fill out questionnaires. For a subset of the respondents, one may wish to evaluate the reliability of the information by conducting follow-up studies, e.g., by conducting detailed interviews or checking official records. The information obtained in the follow-up studies can be used to adjust the estimates of proportions obtained from the questionnaires.

A number of complications can arise in double sampling studies. First, data may accumulate over time or space such that the underlying population structures may be heterogeneous. Second, the verified (cross-classified) subsample may not be a simple random sample of the observations on the covariate. Third, the methods of observation or the covariate that is observed may change over time; at some times or in some places more than one covariate may be observed. We begin by showing that there are two approaches in the literature for interpreting the results from the survey of the covariate (i.e, the error-prone survey) that differ in the way the classification probabilities are defined. We combine the two ap-

		error-prone, J				
		1	2	...	$\mathcal{J}$	
	1	n_{11}	n_{12}	...	$n_{1\mathcal{J}}$	$n_{1.}$
	2	n_{21}	n_{22}	...	$n_{2\mathcal{J}}$	$n_{2.}$
true, I	⋮	⋮	⋮	⋱	⋮	⋮
	$\mathcal{I}$	$n_{\mathcal{I}1}$	$n_{\mathcal{I}2}$	...	$n_{\mathcal{I}\mathcal{J}}$	$n_{I.}$
		$n_{.1}$	$n_{.2}$	...	$n_{.\mathcal{J}}$	n

error-prone, J				
1	2	...	$\mathcal{J}$	
y_1	y_2	...	$y_{\mathcal{J}}$	N

Fig. 1. Notation for the results of testing n units with both an error-prone and an exact test (top) and for the results of testing N units with just the error-prone method (bottom)

proaches in a general model in Section 2 and show in Section 3 that the model allows for partial verification. A repeated disease survey is considered as an example in Section 4. In Section 5 the approach is generalized to consider several test types. The discussion in Section 6 contrasts our approach with the use of hierarchical loglinear models.

We illustrate the logic of the two basic approaches by considering the simple case in which n units are examined using an accurate and an error-prone test; the accurate test assigns an integer value I ranging from 1 to $\mathcal{I}$ to each unit while the error-prone test assigns an integer J ranging from 1 to $\mathcal{J}$. The result is an $\mathcal{I} \times \mathcal{J}$ cross-classified table as in Figure 1. There is also a sample of size N on which observations are made using only the error-prone test. We consider two procedures for selecting these samples. In this paper, the symbols i and j always refer to a realization of the random variables I and J, respectively.

1.1 *Approach 1 – Single population model*

Assume that the n cross-classified units and the N units examined with just the error-prone test are simple random samples of fixed size from the same population. Then, the probability $\mathrm{P}(i \mid j)$ that a unit is actually of type i given that it is classified type j by the error-prone test is the same for both samples. One can estimate these conditional probabilities by

$$\hat{\boldsymbol{P}}(i \mid j) = \hat{\boldsymbol{q}}_{ij} = n_{ij}/n_{\cdot j}$$

where the ^ symbol denotes an estimate and the rest of the notation is as in Figure 1. Denote the $\mathcal{I} \times \mathcal{J}$ matrix with elements $\hat{\boldsymbol{q}}_{ij}$ by $\boldsymbol{Q}$ and the observed vector of error-prone proportions by

$$\boldsymbol{E} = [\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \ldots, \hat{\boldsymbol{e}}_{\mathcal{J}}]^T = \left[\frac{y_1 + n_{\cdot 1}}{N+n}, \frac{y_2 + n_{\cdot 2}}{N+n}, \ldots, \frac{y_J + n_{\cdot \mathcal{J}}}{N+n}\right]^T$$

where, again, the notation is as defined in Figure 1. Let the actual population composition (proportions) be denoted by the vector $\boldsymbol{A} = [a_1, a_2, \ldots, a_{\mathcal{I}}]^T$. Then, intuitively, the actual composition can be estimated by

$$\hat{\boldsymbol{A}} = \boldsymbol{QE}\,.$$

In this example, the conditional probabilities are estimated from the first sample and the marginal probabilities $\mathrm{P}(j)$ are estimated from both samples as $\hat{\boldsymbol{e}}_j$.

This estimator, an example of double sampling, can be shown to be a maximum likelihood estimator (see TENENBEIN, 1970; HOCHBERG, 1977; JOLAYEMA, 1990). It is also an example of stratified random sampling where the units are post-stratified by the results of the error-prone test (SWENSEN, 1988). Related estimators, which are not fully efficient, are discussed by WHITE and CASTLEMAN (1981) and HAND (1986).

1.2 *Approach 2 – Two-population model*

Consider now that the cross-classified sample and the error-prone sample are simple random samples of fixed size obtained from different populations. For example, in the first year of a study both the error-prone and accurate methods might be used but in the second year only the error-prone method is used. The conditional probabilities $P(i \mid j)$ from the first year will not be applicable to the results in the second year if the population composition has changed. To see this, consider the probability that a unit is actually type 1 given that the error-prone test indicates the animal is type 1. If, in the first year, all units are type 1 then all of the units indicated by the error-prone test to be type 1 will in fact be type 1 ($P(I = 1 \mid J = 1) = 1$). If, in the second year, none of the animals are type 1 then none of the units indicated by the error-prone test to be type 1 will in fact be type 1 ($P(I = 1 \mid J = 1)$ is now 0).

There may be a way out of this dilemma. The probability that a unit is classified type j by the error-prone test given that it is actually type i, $P(j \mid i)$, might not vary with the population composition. For example, consider that the covariate is the length of a fish and the actual variable of interest is the age of the fish. The age composition of the fish population will change each year because the number of young fish recruited into the population is highly variable and thus the probability that a fish is a certain age given its size, $P(\text{age} = i \mid \text{length} = j)$, will vary from year to year. But, the distribution of size about age, $P(\text{length} = j \mid \text{age} = i)$, should not change as the population changes in age composition except inasmuch as the growth may be somewhat dependent on environmental conditions. It thus may be entirely reasonable to suppose that $P(\text{length} = j \mid \text{age} = i)$ is constant over time. Similarly, the specificity and sensitivity of a medical test might not vary with the prevalence of the disease (ROGAN and GLADEN, 1978). This assumption would have to be investigated for each application.

The conditional probabilities can be estimated by

$$\hat{\boldsymbol{P}}(j \mid i) = \hat{r}_{ij} = n_{ij}/n_{i\cdot}.$$

where the n_{ij} are the cell counts from a cross-classified sample of size n taken at some time t. Let $\boldsymbol{R}$ denote the $\mathcal{I} \times \mathcal{J}$ matrix with elements $\hat{r}_{ij}$. Also, define the vector $\boldsymbol{E}^*$ to have elements

$$\boldsymbol{E}^* = [y_1/N,\, y_2/N, \ldots, y_{\mathcal{J}}/N]^T,$$

that is, the vector $\boldsymbol{E}^*$ contains estimates of the marginal probabilities $P(j)$ obtained from just the error-prone sample taken at a time t' ($t' \neq t$). Then, the actual composition (proportions) at time t' can be estimated from the moment estimator equations

$$\boldsymbol{E}^* = \boldsymbol{R}^T \hat{\boldsymbol{A}}\,.$$

Thus, assuming $\boldsymbol{R}$ is nonsingular, the actual composition can be estimated by

$$\hat{\boldsymbol{A}} = (\boldsymbol{R}^T)^{-1}\boldsymbol{E}^* \tag{1a}$$

or, more generally, by least squares

$$\hat{\boldsymbol{A}} = (\boldsymbol{R}\boldsymbol{R}^T)^{-1}\boldsymbol{R}\boldsymbol{E}^* \tag{1b}$$

assuming the number of levels of the covariate is greater than or equal to the number of true classes, and assuming $\boldsymbol{R}^T$ is of full column rank.

It can be seen that when the estimates from (1a) are feasible, they are maximum likelihood estimates for Poisson and multinomial data. CLARK (1981) developed an alternative fitting procedure to (1b) which restricts the parameter estimates to the feasible region. HOENIG and HEISEY (1987) developed a model with a more realistic error structure in which the uncertainty in both the classification rates and the distribution of the covariate is accounted for explicitly as functions of the sample sizes.

This approach has appeared in the applied literature as a way to correct deer age composition estimates (SEARLE, 1966 pp. 93–94); correct stock composition estimates for mixed fisheries (WORLUND and FREDIN (1962), FUKUHARA et al. (1962), BERGGREN and LIEBERMAN (1978), PELLA and ROBERTSON (1978), VAN WINKLE et al. (1988)); estimate prevalence of diseases (ROGAN and GLADEN (1978), GREENLAND and KLEINBAUM (1983), HAND (1986)); correct for misclassification in a fourfold table relating disease status to risk factors (KLEINBAUM et al. (1982) and references therein); correct estimates of deer harvest composition obtained from hunter reports (D. Ingebrigtsen, MN Department of Natural Resources, pers. comm.); and convert length-frequency distributions to age-frequency distributions (CLARK (1981), BARTOO and PARKER (1983), KIMURA and CHIKUNI (1987), HOENIG and HEISEY (1987)).

Thus, there are two approaches to using estimates of classification probabilities in conjunction with a vector of observations on a covariate. Method 1 is straight forward, well known, and requires that the classification probabilities be estimated from a sample of the population to which they will be applied. For method 2, the classification rates are conditional on the actual identity rather than on the error-prone identity. The method has been repeatedly derived in the applied literature but does not appear to be well established in the statistical literature. In the next section, we show how the two methods can be combined in a likelihood framework.

2. Combining the Approaches

2.1 *Two surveys*

Assume that we conduct a survey at two times and obtain three samples of fixed size. Sample 1 is a simple random sample of size n_1 collected during the first survey. All n_1 units are examined by the accurate method of classification and the error-prone or surrogate classification method. Samples 2 and 3 are simple random

samples from the second survey and represent the population of interest. All n_2 units in sample 2 are classified by both methods. All N_2 units from sample 3 are classified according to just the error-prone (or surrogate) classification method. The subscript denotes the population (i.e., survey) from which the sample was drawn. We denote the count of units with accurate classification i and surrogate classification j in samples 1 and 2 by n_{ij1} and n_{ij2}, respectively. The count of units in sample 3 with surrogate classification j is denoted by y_{j2}. Again, I ranges from 1 to $\mathcal{I}$ and j from 1 to $\mathcal{J}$. $\mathcal{J}$ must be greater than or equal to $\mathcal{I}$. We assume that $\mathrm{P}(j \mid i)$ for sample 1 is the same as for samples 2 and 3, and we denote this by $\mathrm{P}(j \mid i)_{12}$. In general, subscripts on probabilities are used to denote the population or populations to which the probabilities apply.

Likelihood for Approach 1:

Approach 1 utilizes the information in samples 2 and 3. The likelihood for samples 2 and 3 is the product of independent multinomials and can be written (LITTLE and RUBIN 1987)

$$\Lambda_1 \propto \prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} [\mathrm{P}(i \mid j)_2 \, \mathrm{P}(j)_2]^{n_{ij2}} \prod_{j=1}^{\mathcal{J}} \mathrm{P}(j)_2^{y_{j2}} .$$

There are $\mathcal{I}\mathcal{J} - 1$ parameters to be estimated: $\mathcal{J}(\mathcal{I} - 1)$ conditional probabilities and $\mathcal{J} - 1$ marginal probabilities $\mathrm{P}(j)$. The goal is to estimate the proportion $\mathrm{P}(i)_2$ that is actually in class i and, by the invariance principle of maximum likelihood estimation, this can be accomplished by

$$\hat{\boldsymbol{P}}(i)_2 = \sum_{j=1}^{J} \hat{\boldsymbol{P}}(i \mid j)_2 \, \hat{\boldsymbol{P}}(j)_2 .$$

Likelihood for Approach 2:

Approach 2 utilizes the information in samples 1 and 3. The likelihood is again the product of multinomials

$$\Lambda_2 \propto \prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} [\mathrm{P}(j \mid i)_{12} \, \mathrm{P}(i)_1]^{n_{ij1}} \prod_{j=1}^{\mathcal{J}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)_{12} \, \mathrm{P}(i)_2 \right]^{y_{j2}} .$$

Here, $\mathcal{J}$ must be greater than or equal to $\mathcal{I}$. There are $\mathcal{I}\mathcal{J} + \mathcal{I} - 2$ parameters to be estimated: $\mathcal{I} - 1$ estimates of $\mathrm{P}(i)_1$, $\mathcal{I} - 1$ estimates of $\mathrm{P}(i)_2$, and $\mathcal{I}(\mathcal{J} - 1)$ conditional probabilities.

combined likelihoods:

The likelihood for all of the data can be written as

$$\Lambda_3 \propto \prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} [\mathrm{P}(j \mid i)_{12} \, \mathrm{P}(i)_1]^{n_{ij1}} \prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} [\mathrm{P}(j \mid i)_{12} \, \mathrm{P}(i)_2]^{n_{ij2}} \times \prod_{j=1}^{\mathcal{J}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)_{12} \, \mathrm{P}(i)_2 \right]^{y_{j2}} .$$

We have rewritten the likelihood for samples 2 and 3 in terms of $\mathrm{P}(j \mid i)_{12}$. However, it should be noted that this likelihood reduces to Λ_1 when there are no data from a previous time $(n_1 = 0)$.

2.2 $\mathcal{K} > 2$ *surveys*

The likelihood can be generalized to allow for $\mathcal{K}$ surveys and up to three kinds of samples within a survey. Here, a survey refers to a group of units examined from the same time and place. The survey may produce any and all of the following types of samples: 1) both classification variables are noted on a random sample, 2) just the error-prone or surrogate classifier is noted, and 3) just the accurate classifier is noted. Denote the number classified as type j in the kth survey for a sample in which just the surrogate classifier is used by y_{jk}, the number classified as type i in the kth survey for a sample in which just the accurate classifier is used by x_{ik}, and the number classified as i, j by n_{ijk}. (Note that y_{jk}, x_{ik} and/or n_{ijk} can be 0). Then, the general form of the likelihood is proportional to

$$\Lambda_g \propto \prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} [\mathrm{P}(j \mid i)\, \mathrm{P}(i)_k]^{n_{ijk}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\, \mathrm{P}(i)_k \right]^{y_{jk}} \prod_{i=1}^{\mathcal{I}} \prod_{k=1}^{\mathcal{K}} \mathrm{P}(i)_k^{x_{ik}} . \tag{2}$$

Here, $\mathrm{P}(j \mid i)$ is assumed to pertain to all samples and $\mathrm{P}(i)_k$ pertains to all samples within the kth survey. The generalization to allow for samples for which just the accurate classifier is used has been termed triple sampling by HOCHBERG and TENENBEIN (1983) in the context of Approach 1.

The question of which parameters are estimable may require some thought. Suppose that in the first year of a study just the error-prone technique is used and that no sample cross-classified by the error-prone and accurate tests is obtained in subsequent years. The year one population composition can still be estimated if in at least h subsequent years with different population compositions samples of both type 2 and 3 are obtained, where h is the number of independent conditional probabilities $\mathrm{P}(j \mid i)$ that must be estimated. For example, in a disease survey there are two conditional probabilities (e.g., sensitivity and specificity) that must be estimated. Samples of type 2 and type 3 obtained in years 2 and 3 of a study provide the following system of equations:

$$\hat{e}_{21} = \mathrm{P}(j = 1 \mid i = 1)\, \hat{a}_{21} + \mathrm{P}(j = 1 \mid i = 2)\, \hat{a}_{22} ,$$

$$\hat{e}_{31} = \mathrm{P}(j = 1 \mid i = 1)\, \hat{a}_{31} + \mathrm{P}(j = 1 \mid i = 2)\, \hat{a}_{32} ,$$

where the subscripts x, y refer to year x and category y. The estimates e_{xy} and a_{xy} can be obtained from the samples, and the system of equations can be solved for the conditional probabilties provided the disease prevalence varies across the years.

3. Allowing for Partial Verification

Until now we have assumed that, when a cross-classified table is obtained, the units that are examined by the exact test are a simple random sample of the units for which the covariate (e.g., error-prone test result) was observed. In practice, one is likely to consider the results of the error-prone test or covariate in selecting units for testing with the exact test, e.g., one might use the exact test on equal numbers of apparently-diseased and apparently-disease-free units as determined by the error-prone test. This is known as partial verification and failure to take the stratification into consideration results in what has been termed verification bias (e.g., ZHOU, 1996). There is good reason to consider the results of the error-prone test in selecting the sample for additional testing: otherwise one might obtain by chance a sample in which none of the units were classified as, say, diseased by the error-prone test and one would not be able to estimate some of the classification rates. HAITOVSKY and RAPP (1992) modified Approach 1 to allow for fixed numbers from each error-prone category to be tested with the accurate method.

Here, we show how the general model (2) can be modified to allow for stratification by the covariate. Assume that at a previous time a sample of N_1 units was randomly selected and tested with the error-prone procedure (N_1 fixed) resulting in y^*_{j1} units being classified as type j, $\sum_{j=1}^{\mathcal{J}} y^*_{j1} = N_1$. Suppose further it is decided to use the exact test on $n_{\cdot j1}$ units classified by the error-prone test as type j, for $j = 1, \ldots, \mathcal{J}$, such that $n_1 = \Sigma_j n_{\cdot j1}$ units in all are tested. This results in a cross-classified table with fixed column totals of $n_{\cdot j1}$ and table entries of $n_{\cdot ij1}$.

The likelihood for the N_1 units classified by just the error-prone test is simply a multinomial

$$\Lambda_{N_1} \propto \prod_{j=1}^{\mathcal{J}} \mathrm{P}(j)_1^{y^*_{j1}} = \prod_{j=1}^{\mathcal{J}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\, \mathrm{P}(i)_1 \right]^{y^*_{j1}}.$$

The likelihood for the cross-classified table is the product of $\mathcal{J}$ multinomials, one for each column

$$\Lambda_{n_1} \propto \prod_{j=1}^{\mathcal{J}} \prod_{i=1}^{\mathcal{I}} \mathrm{P}(i \mid j)_1^{n_{ij1}}.$$

Now, in the population at large,

$$\mathrm{P}(i \mid j)_1 = \frac{\mathrm{P}(j \mid i)\, \mathrm{P}(i)_1}{\sum_h \mathrm{P}(j \mid I = h)\, \mathrm{P}(I = h)_1}$$

by Bayes rule. Substituting this into the product of the likelihoods Λ_{N_1} and Λ_{n_1} yields a likelihood in $\mathcal{I}\mathcal{J} - 1$ unknowns: $\mathcal{I}(\mathcal{J} - 1)$ conditional probabilities $\mathrm{P}(j \mid i)$ and $\mathcal{I} - 1$ marginal probabilities $\mathrm{P}(i)_1$. Thus, when the previous data have

been stratified by the results of the error-prone tests, the likelihood still contains information on the conditional probabilities that are used to model the current data. Note that to use approach 2, with stratification by the covariate, it is necessary to know the results of the survey with the error-prone test, i.e., the y^*_{j1}.

The data for the current survey are handled in the same way when the cross-classified table is generated by fixing the numbers in each error-prone category. If $\mathcal{K}$ surveys are conducted, each giving rise to an error prone vector and a cross-classified table obtained by stratifiying on the error-prone test, the complete likelihood for all of the data is proportional to

$$\lambda \propto \prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} \left[\frac{\mathrm{P}(j \mid i)\, \mathrm{P}(i)_k}{\sum_{h=1}^{\mathcal{I}} \mathrm{P}(j \mid I = h)\, \mathrm{P}(I = h)_k} \right]^{n_{ijk}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\, \mathrm{P}(i)_k \right]^{y^*_{jk}}$$

$$= \prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} [\mathrm{P}(j \mid i)\, \mathrm{P}(i)_k]^{n_{ijk}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\, \mathrm{P}(i)_k \right]^{y_{jk}} \tag{3}$$

where $y_{jk} = y^*_{jk} - n_{\cdot jk}$ is the number of observations made using just the error-prone technique. Equation (3) is the same as Equation (2).

Note that the estimator of P (i) for Tenenbein's model (Approach 1) is the same as the estimator based on stratification derived above for the case where the cross-classified data come from the same population as the sample tested with just the error-prone test.

4. Example – A Repeated Disease Survey

We consider a portion of a repeated survey described by NEDELMAN (1988) in which the proportion of people with malaria was estimated in part of Nigeria. Blood smears from some subjects were examined by both a senior and a junior investigator while smears from others were examined by only a junior investigator. Nedelman described some technical aspects of the testing which call into question the use of simple double sampling methodology. Therefore, we use these data only for illustrative purposes and assume the senior investigator to be infallible while the junior investigator is error-prone. A portion of the data is shown in Table 1. In the most recent survey shown (survey 5), only 34 subjects were tested by the infallible classifier. Thus, it is natural to try to make use of the available previously collected information. The conditional proportions in the five surveys were grossly similar: the proportion classified as type 2 by the fallible classifier when the accurate classification was type 2 (i.e., the estimated sensitivity) varied from 84 to 100% with no trend over time; similarly, the proportion classified as type 2 by the fallible classifier when the accurate classification was type 1 (i.e., the complement of the estimated specificity) varied from 0 to 12% with no tem-

Table 1

Malaria survey data from NEDELMAN (1988) pertaining to age class 1. I refers to the true condition and J refers to the error-prone classification (1 = not diseased, 2 = diseased). The conditional probability $\hat{P}(2 \mid i)$ is estimated from the corresponding row of the cross-classification

survey	cross-classification			$\hat{P}(2 \mid i)$	error-prone vector	
	I	$J = 1$	$J = 2$		$J = 1$	$J = 2$
1	1	5	0	0.00	52	173
	2	0	15	1.00		
2	1	7	0	0.00	68	160
	2	3	24	0.89		
3	1	13	2	0.13	90	145
	2	3	16	0.84		
4	1	14	2	0.12	131	157
	2	1	7	0.88		
5	1	10	1	0.09	81	279
	2	1	22	0.96		
1–5	1	49	5	0.09		
	2	8	84	0.91		

poral trend (Table 1). The sample sizes were small so there is little evidence that the conditional classification rates varied among surveys.

Using the method of TENENBEIN (1970) and JOLAYEMA (1990) (Approach 1) on the data from survey 5, the prevalence of malaria is estimated to be 0.754 (line 1, Table 2). The estimate is hardly changed when the cross-classified table from sur-

Table 2

Estimates, $\hat{P}(I = 2)$, of the prevalence of malaria in Nigeria at the time survey 5 was conducted based on the data in Table 1. First line is based on analysis of the cross-classified table and error-prone vector from survey 5; second line, the cross-classified tables from surveys 4 and 5 and the error-prone vector from survey 5; third line, tables and vectors from surveys 4 and 5; fourth line, tables from surveys 1 through 5 and vector from survey 5. In each case, prevalence in the survey year(s) and sensitivity and specificity were estimated. Standard errors are based on the square root of the inverse of the expected information. A small number (0.01) was added to the zeros for surveys 1 and 2 in Table 1 prior to the analysis.

data	standard			number of	
	$\hat{P}(I = 2)$	error	deviance	parameters	df
2 samples, survey 5	0.754	0.043	1.57	3	1
3 samples, surveys 4 & 5	0.767	0.039	2.72	4	3
4 samples, surveys 4 & 5	0.752	0.040	4.75	4	4
6 samples, surveys 1 to 5	0.795	0.034	10.65	7	9

vey 4 is included in the analysis ($\hat{P}(I = 2) = 0.767$; line 2, Table 2) and when the cross-classified table and error-prone vector from survey 4 are included ($\hat{P}(I = 2) = 0.752$; line 3, Table 2). There is a modest reduction in standard error when the data from survey 4 are included. When data from surveys 1 through 5 are used to estimate the year-specific prevalences and the sensitivity and specificity the estimate for survey 5 is $\hat{P}(I = 2) = 0.795$. The standard error of the estimated prevalence for survey 5 is reduced by 21 percent when all five surveys are analyzed instead of just the data from survey 5.

For Tenenbein's estimator (using just survey 5) an exact solution was available. For the other analyses, the likelihood can be numerically maximized using a Newton-Raphson (PRESS et al., 1992), Fisher scoring, or EM algorithm. We used Fisher scoring implemented as iteratively reweighted least squares in SAS Proc NLIN (JENNRICH and MOORE, 1975). Our convergence criterion was that the maximum relative absolute change in the parameters was less than 10^{-8}. We encountered no computational difficulties.

5. Extension to Several Test Types

In general, surveys may be repeated year after year and the observational methods (tests) may evolve over time. This means that several tests may be used in any year. Here, we demonstrate how this can be handled by considering three tests (say, accurate, error-prone reference, and error-prone new) with outcomes I, J, and H, respectively. In any year, up to seven types of samples could be obtained by applying the tests in various combinations (Table 3).

When all samples are simple random samples, the likelihood in (2) can be generalized to

$$\Lambda_{\mathrm{g}} \propto \prod_{i=1}^{\mathcal{I}} \prod_{k=1}^{\mathcal{K}} \mathrm{P}(i)_k^{x_{ik}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\, \mathrm{P}(i)_k \right]^{y_{jk}} \prod_{h=1}^{\mathcal{H}} \prod_{k=1}^{\mathcal{K}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(h \mid i)\, \mathrm{P}(i)_k \right]^{z_{hk}}$$
$$\prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} [\mathrm{P}(j \mid i)\, \mathrm{P}(i)_k]^{n_{ijk}} \prod_{i=1}^{\mathcal{I}} \prod_{h=1}^{\mathcal{H}} \prod_{k=1}^{\mathcal{K}} [\mathrm{P}(h \mid i)\, \mathrm{P}(i)_k]^{m_{ihk}} \prod_{j=1}^{\mathcal{J}} \prod_{h=1}^{\mathcal{H}} \prod_{k=1}^{\mathcal{K}} \mathrm{P}(h, j)_k^{l_{jhk}}$$
$$\prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} \prod_{h=1}^{\mathcal{H}} \prod_{k=1}^{\mathcal{K}} \mathrm{P}(i, j, h)_k^{u_{ijhk}}, \tag{4}$$

where any of the exponents can be zero. It remains to find parameterizations for $\mathrm{P}(h, j)_k$ and $\mathrm{P}(i, j, h)_k$ that introduce as few parameters as possible while maintaining realistic assumptions. Following BAKER (1991), we can make the assumption of independence of the outcomes of the reference and new tests conditional on the trueclassification. Then $\mathrm{P}(h, j)_k$ can be replaced by $\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\ \mathrm{P}(h \mid i)\ \mathrm{P}(i)_k$ and $\mathrm{P}(i, j, h)_k$ can be replaced by $\mathrm{P}(j \mid i)\ \mathrm{P}(h \mid i)\ \mathrm{P}(i)_k$. The resulting model

Table 3

The seven types of samples that may be obtained when three tests are available in survey k ($k = 1, \ldots, \mathcal{K}$). The subscript i refers to the accurate classification ($i = 1, \ldots, \mathcal{I}$); j, to the reference test ($j = 1, \ldots, \mathcal{J}$); h, to the new test ($h = 1, \ldots, \mathcal{H}$)

test(s) used	variable(s) observed	observations
accurate	I	x_{ik}
reference	J	y_{jk}
new	H	z_{hk}
accurate + reference	I, J	n_{ijk}
accurate + new	I, H	m_{ihk}
reference + new	J, H	l_{jhk}
accurate + reference + new	I, J, H	u_{ijhk}

has $\mathcal{I}(\mathcal{K} + \mathcal{J} + \mathcal{H} - 2) - \mathcal{K}$ parameters: $(\mathcal{I} - 1)\,\mathcal{K}$ prevalences $\mathrm{P}(i)_k$, $(\mathcal{K} - 1)\,\mathcal{I}$ conditional probabilities $\mathrm{P}(j \mid i)$, and $(\mathcal{K} - 1)\,\mathcal{I}$ conditional probabilities $\mathrm{P}(h \mid i)$. Alternative formulations, and formulations involving other covariates, are given by BAKER (1991) and are not repeated here.

If stratification or partial verification is used, Equation (4) can be modified in like manner. Assume that when the accurate test is used in combination with either the reference or the new test the stratification is according to the value of the error-prone test (thus, the $n_{\cdot jk}$ and $m_{\cdot hk}$ are fixed rather than random). The counts l_{jhk}, obtained by using the reference and new tests together, could arise by stratifying by the value of either test. The likelihood is

$$\Lambda_{\mathrm{g}} \propto \prod_{i=1}^{\mathcal{I}} \prod_{k=1}^{\mathcal{K}} \mathrm{P}(i)_k^{x_{ik}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\, \mathrm{P}(i)_k \right]^{y_{jk}} \prod_{h=1}^{\mathcal{H}} \prod_{k=1}^{\mathcal{K}} \left[\sum_{i=1}^{\mathcal{I}} \mathrm{P}(h \mid i)\, \mathrm{P}(i)_k \right]^{z_{hk}},$$

$$\prod_{i=1}^{\mathcal{I}} \prod_{j=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{K}} \left[\frac{\mathrm{P}(j \mid i)\, \mathrm{P}(i)_k}{\sum_{t=1}^{\mathcal{I}} \mathrm{P}(j \mid I = t)\, \mathrm{P}(I = t)_k} \right]^{n_{ijk}} \prod_{i=1}^{\mathcal{I}} \prod_{h=1}^{\mathcal{H}} \prod_{k=1}^{\mathcal{K}} \left[\frac{\mathrm{P}(h \mid i)\, \mathrm{P}(i)_k}{\sum_{t=1}^{\mathcal{I}} \mathrm{P}(h \mid I = t)\, \mathrm{P}(I = t)_k} \right]^{m_{ihk}}$$

$$\times \prod_{j=1}^{\mathcal{I}} \prod_{h=1}^{\mathcal{H}} \prod_{k=1}^{\mathcal{K}} \beta(h, j)_k^{\,l_{jhk}} \prod_{j=1}^{\mathcal{I}} \prod_{h=1}^{\mathcal{J}} \prod_{k=1}^{\mathcal{H}} \prod_{k=1}^{\mathcal{K}} \gamma(i, j, h)_k^{\,u_{ijhk}},$$

where the $\beta(h, j)_k$ and $\gamma(i, j, h)_k$ are cell probabilities that depend on the nature of the stratification. If observations are stratified according to the result of the new test when the new and reference tests are used together, the cell probabilities would be

$$\alpha(h, j)_k = \mathrm{P}(j \mid h)_k = \frac{\mathrm{P}(h, j)_k}{\mathrm{P}(h)} = \frac{\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\, \mathrm{P}(h \mid i)\, \mathrm{P}(i)_k}{\sum_{i=1}^{\mathcal{I}} \mathrm{P}(h \mid i)\, \mathrm{P}(i)_k}.$$

If the observations are stratified according to the result of the reference test the cell probabilities would be

$$\alpha(h, j)_k = \mathrm{P}(h \mid j)_k = \frac{\mathrm{P}(h, j)_k}{\mathrm{P}(j)} = \frac{\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\, \mathrm{P}(h \mid i)\, \mathrm{P}(i)_k}{\sum_{i=1}^{\mathcal{I}} \mathrm{P}(j \mid i)\, \mathrm{P}(i)_k} .$$

The cell probabilities $\gamma(i, j, h)_k$ depend on the the nature of the stratification and can be determined the same way as used above. Stratification requires the estimation of no new parameters.

When a new test is introduced, it is not sufficient to obtain a sample cross-classified by the new and reference tests at one time and another sample cross-classified by the reference and accurate tests at another time. Additional information must be available such as:

1) the assumption is made that the results of the new and reference tests are independent conditional on the true (accurate) classification (BAKER 1991)
2) additional covariates are observed on the units (BAKER 1991)
3) surveys are conducted at different times or places with varying prevalences (WALTER and IRWIG, 1988; HUI and WALTER, 1980). In this case, it is not even necessary to have observations on the accurate classification.

6. Discussion

The methods considered here combine two heretofore unrelated approaches. The general approach allows one to use previous information whose cost is essentially free. Often, the previous information on true identity will have been collected according to a scheme with partial verification. This presents no problem if the estimate of the population composition according to the surrogate classifier is estimated for at least one survey in which a cross-classified sample is obtained (so that the $\mathrm{P}(j \mid i)$ are estimable).

ESPELAND and HUI (1987) presented a general method for estimating population proportions when there are several traits of interest and each of these may be observed with accurate and, optionally, error-prone classifying devices. They point out that samples cross-classified by both the accurate and error-prone methods can be from the same population as, or from a different population than, the sample from the population of interest which was only observed with the error-prone method. Thus, their method can accommodate either one of the basic approaches discussed in Section 1. It is instructive to observe what happens if one tries to use their approach for the case where classification rates can be estimated from both prior and current samples (the combined approach of this paper) since Espeland and Hui do not discuss this case.

ESPELAND and HUI (1987) make use of the fact that loglinear models can be used to describe the miscategorization underlying a variety of types of data; they restrict themselves to hierarchical loglinear models, however. Consider a $2 \times 2 \times 3$ contingency table in which the first dimension represents the results of using an accurate classifier (I), the second dimension represents the results of using the error-prone classifier (J), and the third dimension represents the sample (S). The first sample is from population 1, the second and third samples are from population 2. Normally, the data for sample 3 would be incomplete (we only observe the marginal frequencies for the error-prone classifier). However, for the purpose of studying the structure of the problem we assume sample 3 is completely observed. The probability P (error-prone results $= j \mid$ accurate results $= i$) is the same for all three samples; in addition, the probabilities $P(J = j)$ and $P(I = i)$ do not vary among the samples 2 and 3. The loglinear model log (count) $= I^*J\ I^*S$ fits samples 1 and 2 (or samples 1 and 3) – this model corresponds to approach 2. The same model fits samples 2 and 3 but, additionally, the more restricted model log (count) $= I^*J\ S$ (corresponding to approach 1) fits these samples. We need to include both two-factor interactions if we wish to fit a hierarchical loglinear model to all three samples. Thus, the additional information about samples 2 and 3 cannot be included in the model if sample 1 is included in the data. Use of the model with two two-factor interactions is not fully efficient.

The use of data from a variety of populations is based on the assumption that the classification rates (P(error | actual)) have not changed from sample to sample. This assumption is commonly made in medical applications and can be tested using standard methods such as a likelihood ratio test.

Acknowledgements

We thank an anonymous reviewer for helpful comments. This is VIMS contribution No. 2444.

References

BAKER, S. G., 1991: Evaluating a New Test Using a Reference Test with Estimated Sensitivity and Specificity. *Communications in Statistics – Theory and Methods* **20**, 2739–2752.

BARTOO, N. and PARKER, K., 1983: Stochastic Age-Frequency Estimation Using the von Bertalanffy Growth Equation. U.S. National Marine Fisheries Service Fishery Bulletin **81**, 91–96.

BERGGREN, T. J. and LIEBERMAN, J. T., 1978: Relative Contribution of Hudson, Chesapeake, and Roanoke Striped Bass, *Morone saxatilis*, Stocks to the Atlantic Coast Fishery. U.S. National Marine Fisheries Service Fisheries Bulletin **76**, 335–345.

CLARK, W., 1981: Restricted Least-Squares Estimates of Age Composition from Length Composition. *Canadian Journal of Fisheries and Aquatic Science* **38**, 297–307.

ESPELAND, M. A. and HUI, S. L., 1987: A General Approach to Analyzing Epidemiologic Data that Contain Misclassification Errors. *Biometrics* **43**, 1001–1012.

FUKUHARA, F. M., MURAI, S., LALANNE, J. J., and SRIBHIBHADH, A., 1962: Continental Origin of Red Salmon as Determined from Morphological Characters. *International North Pacific Fisheries Commission Bulletin* **8**, 400–408.

GREENLAND, S. and KLEINBAUM, D. G., 1983: Correcting for Misclassification in Two-way and Matched-pair Studies. *International Journal of Epidemiology* **12**, 93–97.

HAITOVSKY, Y. and RAPP, J., 1992: Conditional Resampling for Misclassified Multinomial Data with Applications to Sampling Inspection. *Technometrics* **34**, 473–483.

HAND, D. J., 1986: Estimating Class Sizes by Adjusting Fallible Classifier Results. *Computations & Mathematics with Applications* **12A**, 289–299.

HOCHBERG, Y., 1977: On the Use of Double Sampling Schemes in Analyzing Categorical Data with Misclassification Errors. *Journal of the American Statistical Association* **72**, 914–921.

HOCHBERG, Y. and TENENBEIN, A., 1983: On Triple Sampling Schemes for Estimating from Binomial Data with Misclassification Errors. *Communications in Statistics – Theory and Methods* **12(13)**, 1523–1533.

HOENIG, J. M. and HEISEY, D. M., 1987: Use of a Loglinear Model with the EM Algorithm for Correcting Estimates of Stock Composition and Converting Length to Age. *Transactions of the American Fisheries Society* **116**, 232–243.

HUI, S. L. and WALTER, S. D., 1980: Estimating the Error Rates of Diagnostic Tests. *Biometrics* **36**, 167–171.

JENNRICH, R. I. and MOORE, R. H., 1975: Maximum Likelihood Estimation by Means of Nonlinear Least Squares. *Proceedings of the American Statistical Association Statistical Computing Section* 57–65.

KIMURA, D. and CHIKUNI, S., 1987: Mixtures of Empirical Distributions: an Iterative Application of the Age-Length Key. *Biometrics* **43**, 23–35.

KLEINBAUM, D. G., KUPPER, L. L., and MORGENSTERN, H., 1982: Epidemiologic Research: Principles and Quantitative Methods. Belmont, CA: Wadsworth.

JOLAYEMA, E. T., 1990: Relative Frequency Estimation in Multiple Outcome Measurement with Misclassification. *Biometrical Journal* **6**, 707–711.

LITTLE, R. J. A. and RUBIN, D. B., 1987. Statistical Analysis with Missing Data. New York: Wiley.

NEDELMAN, J., 1988: The Prevalence of Malaria in Garki, Nigeria: Double Sampling with a Fallible Expert. *Biometrics* **44**, 635–655.

PELLA, J. J. and ROBERTSON, T. L., 1978: Assessment of Composition of Stock Mixtures. U.S. National Marine Fisheries Service Fisheries Bulletin **76**, 415–423.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., and FLANNERY, B. P., 1992: Numerical Recipes in FORTRAN the Art of Scientific Computing, 2nd Edition. Cambridge University Press, New York.

ROGAN, W. J. and GLADEN, B., 1978: Estimating Prevalence from the Results of a Screening Test. *American Journal of Epidemiology* **107**, 71–76.

SEARLE, S., 1966: Matrix Algebra for the Biological Sciences (Including Applications in Statistics). Wiley, New York.

SWENSEN, A. R., 1988: Estimating Change in a Proportion by Combining Measurements on a True and a Fallible Classifier. *Scandinavian Journal of Statistics* **15**, 139–145.

TENENBEIN, A., 1970: A Double Sampling Scheme for Estimating from Binomial Data with Misclassification. *Journal of the American Statistical Association* **65**, 1350–1361.

VAN WINKLE, W., KUMAR, K. D., and VAUGHAN, D. S., 1988: Relative Contributions of Hudson River and Chesapeake Bay Striped Bass Stocks to the Atlantic Coastal Population. *American Fisheries Society Monograph* **4**, 255–266.

WALTER, S. D. and IRWIG, L. M., 1988: Estimation of Test Error Rates, Disease Prevalence and Relative Risk from Misclassified Data: a Review. *Journal of Clinical Epidemiology* **41**, 923–937.

WHITE, B. S. and CASTLEMAN, K. R., 1981: Estimating Cell Populations. *Pattern Recognition* **13**, 365–370.

WORLUND, D. D. and FREDIN, R. A., 1962: Differentiation of Stocks. In: Symposium on Pink Salmon, H.R. MacMillan Lectures in Fisheries. University of British Columbia, Vancouver, 143–153.

ZHOU, X. H., 1996: A Nonparametric Maximum Likelihood Estimator for the Receiver Operating Characteristic Curve Area in the Presence of Verification Bias. *Biometrics* **52**, 299–305.

authors' affiliations

JOHN M. HOENIG, Virginia Institute of Marine Science, College of William and Mary, Gloucester Point, VA 23062, U.S.A.

R. CHOUDARY HANUMARA, Department of Computer Science and Statistics, University of Rhode Island, Kingston, RI 02881, U.S.A.

DENNIS M. HEISEY, Department of Surgery, University of Wisconsin, Madison, WI 53705, U.S.A.

JOHN HOENIG
VIMS
P.O. Box 1346
Gloucester Pt.
VA 23062
U.S.A.
E-mail: hoenig@vims.edu

Received, September 2000
Revised, November 2001
Accepted, January 2002