

SIMULATION OF DATA FOR MONTE CARLO STUDIES OF PREVALENCE ESTIMATION IN TWO-PHASE CROSS-SECTIONAL EPIDEMIOLOGIC STUDIES

Mark S. Litaker, Medical College of Georgia, Augusta, GA
 Martin C. Weinrich, University of South Carolina, Columbia, SC

Abstract

Estimators of prevalence in two-phase epidemiologic studies, in which data are available from the results of a screening test for disease, but definitive diagnoses are available only for subsamples of those in each of the screening categories, are typically based on a stratified binomial model, and a weighted average of stratum-specific prevalences. Estimators based on a multinomial model, or on iterative techniques, may provide more precise prevalence estimation in some situations by using the entire sample in the calculation of the point estimate, rather than using only those observations which receive both tests. Simulation studies are a useful tool in evaluation of the bias and stability of alternative estimators, since the true values of prevalence and the sensitivity and specificity of the screening test are unknown in actual practice. A SAS® program is presented which produces data from simulated two-phase studies, and allows the specification of true prevalence, sensitivity, specificity, and the sampling fractions within positive- and negative-screening strata. The program calls macros which calculate the prevalence estimates and their sample bias and variance.

Introduction

Prevalence is defined as the proportion of subjects in a population who have a disease or other characteristic of interest at a point in time. Prevalence may be estimated as the proportion of subjects in a random sample from the population who have the disease, as identified by a diagnostic test. Often in epidemiologic studies, the application of a definitive diagnostic test to all subjects in a large population-based sample is not feasible, due to cost or ethical constraints, since such a test may be expensive or invasive. If the disease is rare, very many tests may be required to identify enough cases to allow precise estimation of the prevalence. For many diseases, relatively inexpensive screening tests are available, which allow preliminary classification of subjects into high-prevalence (positive-screening) and low-prevalence (negative-screening) strata. In such situations, a two-phase study design may be used, in which the screening test is applied to a large sample, and then smaller samples are chosen from the positive- and negative-screening strata. The smaller samples then receive the definitive diagnostic test. The study may be summarized in the following 2x3 table of observed counts:

	Disease	No Disease	Unknown
Positive Screen	n1	n2	n3
Negative Screen	n4	n5	n6

Prevalence Estimators for Two-Phase Studies

Typically, the prevalence estimate from such a study is the weighted average of the prevalences among those receiving the definitive diagnostic technique in the two strata, with the weights being the proportions of the total sample in each of the strata (Tenenbein, 1970). The proportions chosen for the second stage of the study may differ between the strata, and may be chosen to maximize the precision of estimation for a given total sample size or for a given total cost. A potential weakness of this estimator is that the stratum-specific prevalence estimates may be based on relatively few subjects, and thus may be imprecise. In an extreme case, the definitive diagnosis may be applied only to subjects who show positive results on the screening test, and no estimate of the prevalence in the negative-screening stratum is available. In this case, the weighted-average prevalence estimator is not directly applicable.

An estimate of prevalence which is applicable in this situation if the performance characteristics of the screening test are known from previous study, is based on a multinomial model, with category probabilities defined as functions of the unknown true prevalence and the known sensitivity and specificity of the screening test (Weinrich and Litaker, 1996). In this model, information from subjects who receive only the screening test contribute directly to the calculation of the prevalence estimator. A second estimator which may be applied in this situation uses the EM algorithm to iteratively fill in a 2x2 table as if all subjects received the true diagnostic test (Little and Rubin, 1987). This model is derived from considering the two-phase study design as a binomial estimation problem with missing data, and it also requires knowledge of the sensitivity and specificity of the screening test.

Monte Carlo Studies

Simulation, or Monte Carlo, studies are a useful tool in evaluating the relative performance of the various estimators across a large number of replications of studies with known characteristics. First-stage samples are generated from 'populations' with known prevalence and screening tests with known characteristics are applied to these samples. Second-stage samples of varying sizes may be used. The performance of different estimators may then be evaluated in a variety of study scenarios.

Simulation of Two-Phase Studies Using SAS

The accompanying SAS macro program simulates a two-phase study as a random sample from a population with the assigned prevalence, the application of a screening test with known sensitivity and specificity, and an evaluation of samples from the screening strata using a definitive diagnostic procedure. The macro call includes parameters to specify the true prevalence, screening test sensitivity and specificity, the number of observations in a single study replication or table, sampling fractions for each stratum, number of tables to be generated, and a seed value for the random number streams. Simulation of data for individual subjects occurs in three steps, with each step utilizing a stream of random numbers from the uniform distribution, generated by the RANUNI function. In the first step, the true (but unobserved) disease state is defined for each subject, based on the prevalence. A random number between 0 and 1 is assigned to each subject. If this number is less than or equal to the true prevalence, this subject is assigned as 'diseased'. A second uniform random variate is used in the same way to assign screening test results, using the fixed values of sensitivity and specificity. The third random number is used to identify the random subsamples of the positive- and negative-screening strata, which will receive the (observed) definitive diagnostic technique. For subjects selected in the third step, the observed diagnosis is assigned as equal to the true disease status, since the definitive technique is assumed to be without error.

Based on the screening and diagnostic test results, the subject is assigned membership in one of the six table cells. The SUMMARY procedure is used to sum the individual cell counts, and thus generate the observed total numbers of observations in each cell, n1 to n6, for each table. The simulation of data for the required number of repetitions of the study is complete at this point in the program.

The proportion of individuals with the true disease state is calculated, as a check on the simulation procedure, and to compare the efficiency of the two-phase design with that of complete evaluation using only the definitive diagnosis. The estimates of prevalence are calculated, using additional macro calls. The MEANS procedure is used to calculate descriptive statistics which are used to evaluate

the bias and precision of the estimators.

To allow repeated calls of the macro with limited computer storage, the DATASETS procedure is used to clear all temporary data sets between simulation runs.

The Simulation Program

```
/* Generate data for incompletely classified
2-way tables with diagnosis present for
samples of those screening positive and of
those screening negative */
```

```
%macro rp(prev,sens,spec,table_n,f1,f2,
          tables, seed);

data data1;
* set seeds for random number generator;
  seed=&seed;
  seed1=round(seed*5);
  seed2=round(seed*11);
  seed3=round(seed*25);

* set prevalence, sensitivity, and
specificity of screen;
  prev=&prev;
  sens=&sens;
  spec=&spec;

/* set sampling fractions: the proportion
of subjects in each screening stratum to
receive the definitive diagnosis */
  f1=&f1;
  f2=&f2;

* set number of subjects per table;
  table_n=&table_n;

* set number of tables to be simulated;
  tables=&tables;

do i = 1 to tables;
  table = i;
  do j = 1 to table_n;

* generate 3 random number streams ;
    rand1=ranuni(seed1 + i + j);
    rand2=ranuni(seed2 + i + j);
    rand3=ranuni(seed3 + i + j);

/* generate data for each subject
and assign true disease status */

    if rand1 le prev then ds = 1;
    else if rand1 gt prev then ds = 0;

/* assign screening test results */

* subjects with true disease;
    if ds=1 and rand2 le sens then screen = 1;
```

```

else if ds=1 and rand2 gt sens
    then screen = 0;

* non-diseased subjects;
if ds=0 and rand2 le spec then screen = 0;
    else if ds=0 and rand2 gt spec
        then screen = 1;

/*
Assign results of definitive diagnostic
test, and assign missing status to random
samples of positives and negatives, to get
a sample of subjects with known diagnosis
and a sample without known diagnosis
*/
dx=ds;
    *(dx is the observed diagnosis);

* positive screens
    (sampling fraction = f1);
if screen eq 1 and rand3 gt f1 then dx = .;

* negative screens
    (sampling fraction = f2);
if screen eq 0 and rand3 gt f2 then dx = .;
output;
end;
end;
keep table table_n prev sens spec
    dx screen ds;
run;

data data1;
set data1;

/* Identify table cell to which each
subject belongs, based on screening test
and diagnosis or absence of diagnosis */

cell1=0; cell2=0; cell3=0; cell4=0;
cell5=0; cell6=0;

if screen=1 and dx=1 then cell1=1;
    * (pos screen, pos dx);
else if screen=1 and dx=0 then cell2=1;
    * (pos screen, neg dx);
else if screen=1 and dx = . then cell3=1;
    * (pos screen, unk dx);
else if screen=0 and dx=1 then cell4=1;
    * (neg screen, pos dx);
else if screen=0 and dx=0 then cell5=1;
    * (neg screen, neg dx);
else if screen=0 and dx = . then cell6=1;
    * (neg screen, unk dx);

keep prev sens spec table table_n ds
cell1-cell6;
run;

/* sum individual cell memberships
to get cell counts for each table */

```

```

proc summary;
class table table_n prev sens spec;
var cell1-cell6 ds;
output out=tables sum=n1-n6 ds ;
data tables;
set tables;
if _type_ eq 31;

* assign empty cells zero frequencies;
    if n1 = . then n1 = 0;
    if n2 = . then n2 = 0;
    if n3 = . then n2 = 0;
    if n4 = . then n4 = 0;
    if n5 = . then n5 = 0;
    if n6 = . then n6 = 0;

/* Calculate estimates of prevalence
and error of estimates
(difference from true prevalence) */

/* Binomial estimate based on true disease
state with no missing information
(observed prevalence for the complete
table) */

binp=ds/table_n;
vbinp=(ds/table_n)*(1-(ds/table_n))
    /table_n; * (estimated variance);
binp_err=binp-prev; * (error);

/* Maximum-likelihood estimate based on
multinomial model */

%ml(sens,spec,n1,n2,n3,n4,n5,n6);

/* Estimate from EM algorithm */

%em(sens,spec,n1,n2,n3,n4,n5,n6);

/* Weighted-average estimate based
on stratified binomial sample */

%pwt(sens,spec,n1,n2,n3,n4,n5,n6);

keep prev sens spec table_n table n1-n6
binp vbinp binperr;
run;

proc means data=tables n mean var min max;
title2 "prev=&prev, sensitivity=&sens,
specificity=&spec";
title3 "table size= &table_n, number of
tables= &tables";
title4 "sampling fractions: &f1 (pos), &f2
(neg)";
var prev sens spec table_n binp vbinp;
run;
proc datasets kill;
%mend rp;
run;

```

Conclusion

Simulation can be an important tool in the evaluation of the performance of statistical estimators. This program illustrates the use of functions and procedures available in base SAS software to implement simulation of two-phase epidemiologic studies with specified characteristics and summarization of the resulting parameter estimates.

References

Tenenbein A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association* 65, 1350-1361.

Weinrich MC, Litaker MS. (1996). Prevalence Estimation from Screening Data Using Maximum Likelihood. Presentation at International Biometric Society/ENAR 1996 Spring Meeting, March, 1996.

Little RJA, Rubin D. (1987). Statistical Analysis with Missing Data. John Wiley and Sons, NY.

SAS is a registered trademark of SAS Institute, Inc., in the USA and other countries. ® indicates USA registration.

Mark Litaker
Office of Biostatistics
Medical College of Georgia
Augusta, GA 30912-4900
(706) 721-3785
mlitaker@mail.mcg.edu