REPORT ON THE AGRICULTURE SURVEY DATA SIMULATION PROGRAM*

Dawn E. Haines
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203

Institute of Statistics Mimeo Series No. 2286

June 28, 1996

Abstract

In this technical report, we describe a computer program that generates data useful for comparing different estimation methods used to analyze Agriculture Survey data. The generated data is formulated as an intercept model where the intercept represents an interviewer effect. In particular, the simulation study involves adding artificial interviewer error to the Agriculture Survey data. The estimation methods are for the type of data where a primary response variable (y) and a covariate (x) are measured for each sampled unit. The model $y_{ij} = \mu_{ij} + b_i + \epsilon_{ij}$ is utilized where μ_{ij} is the true value of the Agriculture Survey response, b_i is the systematic errror due to the i^{th} interviewer, and ϵ_{ij} is a random, independent measurement error for all $i = 1, \ldots, I$ interviewers and $j = 1, \ldots, m$ units. Thus, the simulated survey response y_{ij} is the sum of the true value, the interviewer variable, and the random error. Additionally, the covariate x_{ij} is defined as $x_{ij} = \mu_{ij} + e_{ij}$, where e_{ij} is a random error term used to link the covariate x_{ij} with the true primary response variable μ_{ij}. The random variables $\{b_i\}$, $\{\epsilon_{ij}\}$, and $\{e_{ij}\}$ are mutually independent.

KEY WORDS: Response Bias, Interviewer Effect, Measurement Error

*This work was supported by a grant from the National Agricultural Statistics Service, United States Department of Agriculture, Contract # 500-95-0038.

1 INTRODUCTION

Consider a population consisting of N units subdivided into l strata. Let N_h denote the size of the h^{th} stratum. For each stratum, I different interviewers are used to collect data. That is, the interviewers are nested within each stratum. Each interviewer in the h^{th} stratum collects data on m units selected randomly without replacement, where m is the integer part of $(N_h \cdot sf1)/I$ and $sf1$ is a user-specified sampling fraction. Thus, a SRSWOR of size $n1h = I \cdot m$ is selected from the h^{th} stratum and $n1h$ is approximately a specified fraction, $sf1$, of the stratum size N_h. Here, m varies with strata.

Data on a primary response variable and a covariate, with possibly some measurement error, are collected on these $n1h$ sampled units. A model for the observed variables in the h^{th} stratum is:

$$\begin{aligned} y_{ij} &= \mu_{ij} + b_i + \epsilon_{ij}, \\ x_{ij} &= \mu_{ij} + e_{ij}, \end{aligned}$$

where

$$\begin{aligned} i &= 1, \ldots, I \\ j &= 1, \ldots, m. \end{aligned}$$

Here μ_{ij} denotes the true value of the primary response variable, b_i is the systematic error due to the i^{th} interviewer, ϵ_{ij} is the respondent measurement error, and e_{ij} is a random error term used to connect the covariate x_{ij} with the true primary response variable μ_{ij}. We assume $b_i \sim \text{NID}(B_b, \sigma_b^2)$, $\epsilon_{ij} \sim \text{NID}(B_\epsilon, \sigma_\epsilon^2)$, and $e_{ij} \sim \text{NID}(0, \sigma_e^2)$. In addition, $\{b_i\}$, $\{\epsilon_{ij}\}$, and $\{e_{ij}\}$ are mutually independent.

To get estimates of the interviewer effects and measurement errors, a subsample of size $n2h$ is selected from the first sample of size $n1h$ in the following way. Let $n2h = k \cdot I + f$. If $f = 0$, then k out of m units are selected at random from each interviewer. If $f > 0$, then f of the interviewers are selected at random first. For these f selected interviewers, $k + 1$ units are selected at random from each of their m units. The remaining $I - f$ interviewers have only k units selected at random from each of their m units. The subsample size $n2h$ is

taken to be the integer part of $n1h \cdot sf2$ where $sf2$ is a prespecified subsampling fraction. Estimators are constructed based on the information from both samples.

The purpose of our program is to generate first stage samples, subsamples, and corresponding response variables. To generate such data, several input variables are required. In our program, we first describe the input variables. We then describe how to generate values necessary for the construction of l strata. Given N and l, the stratum sizes, N_h, $h = 1, \ldots, l$ are first generated according to a specified distribution. Stratum means (μ) of the response variable (μ_r) are generated according to a specified distribution. Stratum variances are then calculated as a function of the stratum means and a prespecified coefficient of variation. The population mean and variance are computed from the stratum sizes, means, and variances. After individual strata variables are constructed, N_h population values of the primary response variables (μ_r) and covariates (x_r) from a specified distribution are generated for each strata. Here, the subscript r is a dummy variable used to denote the r^{th} stratum element, where $r = 1, \ldots, N_h$.

Once the N_h pairs of (μ_r, x_r) are generated for each strata, we select the first phase sample of size $n1h$ units. We assign the $n1h$ units randomly to the I interviewers, m respondents per interviewer. The next step is to generate the observed values of the response variable ,y_{ij}, as the sum of μ_{ij}, b_i, and ϵ_{ij}, where these variables represent the true value of y_{ij}, the systematic error due to the i^{th} interviewer, and a random measurement error term, respectively. Here, μ_{ij} and x_{ij} denote the values of μ_r and x_r, respectively, for the j^{th} respondent of the i^{th} interviewer, where $j = 1, \ldots, m$ and $i = 1, \ldots, I$. Finally, a simple random subsample of size $n2h$ is selected from $n1h$ first stage samples as described above.

2 INPUT VARIABLES

The population size N, number of strata l, number of interviewers I, sampling fraction $sf1$, and subsampling fraction $sf2$ are specified by the user. The primary response variable, μ_r, is assumed to have a symmetric Beta(α, α) distribution on the support (0,2·*mbarh*), where *mbarh* is generated according to a distribution given below. The value of α is also given below. We chose the Beta(α, α) distribution since we are interested in a "normal-shaped" distribution with a nonnegative support.

By noting the properties of a Beta(α, α) distribution on the support (0,2·*mbarh*), we solve for the unknown parameter α. The mean of this distribution, μ, is equal to *mbarh* while the standard deviation is

$$\sigma_\mu = \frac{mbarh}{\sqrt{2\alpha + 1}}.$$

As a result, the population coefficient of variation is

$$cv = \frac{1}{\sqrt{2\alpha + 1}}.$$

Solving for the parameter α yields

$$\alpha = \frac{1}{2}\left(\frac{1}{cv^2} - 1\right).$$

Since the parameter α is strictly > 0, the input variable *cv* is forced to be strictly < 1. The coefficient of variation *cv* is specified by the user and α is computed by the program.

The mean, *mbarh*, of the primary response variable μ_r is also generated according to a Beta(α, α) distribution on the support $(mbarl, mbaru)$, where the lower and upper limits, denoted *mbarl* and *mbaru*, respectively, are specified. The parameter α is given above.

Once *mbarh* is generated, the corresponding stratum variance σ_μ^2, represented as *sig2h*, is computed as

$$\sigma_\mu^2 = (mbarh \cdot cv)^2.$$

The interviewer correlation ρ_y, labelled *rhoy*, is another input variable. This variable measures the extent to which response errors made by respondents of the same interviewer are correlated. The value of *rhoy* is defined as

$$\frac{\sigma_b^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_\epsilon^2}.$$

A customary range for the within interviewer correlation is 0 to .30. In our example, *rhoy* is assigned the value .05.

Another variable of interest is reliability. Reliability is defined as the ratio of true variation to total variation. Thus, reliability represents the proportion of total variation not due to measurement error. Statistically, reliability is defined as

$$\frac{\mathrm{Var}\{\mathrm{E}(y_{ij} \mid (i,j))\}}{\mathrm{Var}(y_{ij})},$$

which has the value

$$\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_\epsilon^2}.$$

A common range for reliability is .40 to 1. Reliability, named *rel* in our program, is fixed at .80. Given ρ_y, rel, and σ_μ^2, we compute σ_b^2 and σ_ϵ^2 as

$$\sigma_b^2 = \frac{sig2h \cdot \rho_y}{rel},$$

and

$$\sigma_\epsilon^2 = \frac{sig2h(1 - rel - \rho_y)}{rel}.$$

Interviewer bias is another variable we consider in the Agriculture Survey. The presence of interviewer bias may lead to differences in the mean values of y_{ij} obtained by different interviewers who are sampling comparable parts of the same population. In our program, the mean interviewer bias B_b is defined as a percentage, *intbias*, of the stratum mean. That is,

$$B_b = intbias \cdot mbarh.$$

We assume no interviewer bias in our program, and thus $intbias = 0$. One additional bias term which may play a role in our study is the mean respondent bias. Respondent bias is primarily concerned with possible differences between the observed and true values of measured characteristics. It is designed to measure any difference that may exist between the original survey response and the reinterview response. In our program, the mean response bias B_ϵ is defined as a percentage, *respbias*, of the stratum mean. *Respbias* is defined as .10. Hence, we represent the mean response bias as

$$B_\epsilon = respbias \cdot mbarh.$$

The next variable in our study is $\rho_{\mu x}$, the correlation between the variables μ_r and x_r. *Rhomux*, the name given to $\rho_{\mu x}$, is assigned the value .50. Given $\rho_{\mu x}$, we compute the variance of the error e_r, σ_e^2, as

$$\sigma_e^2 = \frac{sig2h(1 - \rho_{\mu x})}{\rho_{\mu x}}.$$

3 METHODOLOGY

Given the various input variables, we first generate values necessary for the construction of l strata. For $h = 1, \ldots, l-1$, the stratum sizes, N_h, are generated uniformly on the interval (NL, NU) and are required to be integer-valued. The lower and upper stratum sizes we are willing to accept are designated NL and NU, respectively, and are defined as

$$NL = \frac{N}{l} - \frac{N}{l(l-1)},$$

and

$$NU = \frac{N}{l} + \frac{N}{l(l-1)}.$$

The choice of NL and NU are such that the stratum sizes N_h will be between 1 and N and such that the sum

$$\sum_{h=1}^{l-1} N_h \leq N.$$

We define

$$N_l = N - \sum_{h=1}^{l-1} N_h.$$

Now, N_h is uniformly distributed on (NL, NU), which implies

$$\frac{N_h - NL}{NU - NL} \sim \text{Uniform}(0, 1).$$

Designating the Uniform(0,1) random variable as u and solving for N_h yields

$$N_h = (NU - NL)u + NL.$$

The stratum means $mbarh$ are distributed as symmetric Beta(α, α) random variables on the support $(mbarl, mbaru)$, where

$$\alpha = \frac{1}{2}\left(\frac{1}{cv^2} - 1\right).$$

The value of cv, the coefficient of variation, is specified by the experimenter and is a known function of the mean and standard deviation of μ_r. As a result, we calculate the individual stratum variances, termed $sig2h$, using the formula $sig2h = (mbarh \cdot cv)^2$.

To understand overall population characteristics, strata information is combined to yield the population mean and variance, denoted as $mu0$ and $sigma2$, respectively. The population mean is calculated using the formula

$$mu0 = \sum_{h=1}^{l} \frac{N_h \cdot mbarh}{N}.$$

It is possible to construct the population variance, $sigma2$, if the stratum sizes, means, and variances are known in addition to the population mean. According to Cochran (1977), the population variance, $sigma2$, is calculated as

$$\begin{aligned} sigma2 &= \frac{1}{N}\left(\sum_{h=1}^{l}(N_h-1)(sig2h)+\sum_{h=1}^{l}N_h(mbarh-mu0)^2\right) \\ &= \frac{1}{N}\sum_{h=1}^{l}(N_h-1)(sig2h)+\frac{1}{N}\sum_{h=1}^{l}N_h(mbarh)^2-\frac{1}{N}\sum_{h=1}^{l}N_h(mu0)^2 \\ &= \frac{1}{N}\sum_{h=1}^{l}(N_h-1)(sig2h)+\frac{1}{N}\sum_{h=1}^{l}N_h(mbarh)^2-(mu0)^2 \\ &= \frac{1}{N}\sum_{h=1}^{l}\left((N_h-1)(sig2h)+N_h(mbarh)^2\right)-(mu0)^2. \end{aligned}$$

After individual strata variables are constructed as well as the population mean and variance, we generate N_h population variables for each strata. First, the primary response variable, μ_r, is generated from a Beta(α,α) distribution on the support (0,2·$mbarh$). The resulting mean of μ_r is

$$\mathrm{E}(\mu_r) = \mu = mbarh,$$

while the variance of μ_r is

$$\mathrm{Var}(\mu_r) = \sigma_\mu^2 = \frac{(mbarh)^2}{2\alpha+1}.$$

A random measurement error term, e_r, which is normally distributed with mean 0 and variance

$$\sigma_e^2 = \frac{sig2h(1-\rho_{\mu x})}{\rho_{\mu x}},$$

is constructed. Then, x_r is computed as $\mu_r + e_r$. Note that x_r has mean

$$\mathrm{E}(x_r) = mbarh,$$

and variance

$$\mathrm{Var}(x_r) = \sigma_x^2 = \frac{(mbarh)^2}{2\alpha + 1} + \frac{sig2h(1 - \rho_{\mu x})}{\rho_{\mu x}}.$$

Recall that $\rho_{\mu x}$ is the correlation between the variables μ_r and x_r, and is denoted $rhomux$ in the program. If x_r's value is negative, it is automatically reassigned the value 0. This step ensures that the random variable x_r has a nonnegative support.

Next, we draw the phase 1 SRSWOR of size $n1h$, where $n1h$ is a function of $I, sf1$, and N_h. To draw this sample, we generate a Uniform(0,1) random number called u_r to associate with each value of the dummy variable r. For each strata, we then sort by u_r. This yields a random permutation of the elements within each strata. Our SRSWOR is comprised of the first $n1h$ values in each strata. The *first* $n1h$ observations are chosen simply for convenience.

The next task to consider is constructing the observed y_{ij} values for $i = 1, \ldots, I$ interviewers and $j = 1, \ldots, m$ respondents. Note that the first phase sample size $n1h$ equals the number of interviewers, I, times the interviewer assignment size, m. As previously stated, our model is

$$y_{ij} = \mu_{ij} + b_i + \epsilon_{ij},$$

where b_i and ϵ_{ij} represent the interviewer and respondent measurement error terms, respectively. Remember that μ_r and x_r denote μ_{ij} and x_{ij}, respectively, where the subscripts i and j represent the j^{th} unit of the i^{th} interviewer in the phase 1 sample.

The distribution of b_i is Normal with mean B_b and variance σ_b^2, where

$$B_b = \mathrm{E}(b_i) = intbias \cdot mbarh,$$

and

$$\sigma_b^2 = \mathrm{Var}(b_i) = \frac{sig2h \cdot \rho_y}{rel}.$$

In our program, b_i represents an interviewer variable whose value remains constant for each interviewer but varies among interviewers and strata. The final item to generate is the random error term. Particularly, each interviewer-respondent combination has an associated error term, namely ϵ_{ij}. These errors are normally distributed with mean B_ϵ and variance σ^2_ϵ, where

$$B_\epsilon = \mathrm{E}(\epsilon_{ij}) = respbias \cdot mbarh,$$

and

$$\sigma^2_\epsilon = \mathrm{Var}(\epsilon_{ij}) = \frac{sig2h(1 - rel - \rho_y)}{rel}.$$

Once the three individual pieces of our model are constructed, we sum them together. Furthermore, the Agriculture Survey response y_{ij} is required to be nonnegative. Hence, we take y_{ij} to be the maximum of 0 and the sum $\mu_{ij} + b_i + \epsilon_{ij}$. Thus, all response values are nonnegative. Another detail to address is the inclusion of elements in our phase 1 SRSWOR using the indicator variable, $inds1$. If $inds1 = 1$ for an observation, then that variable is included in the first phase sample. Otherwise, $inds1 = 0$.

Up to this point, we have drawn a SRSWOR of size $n1h$ from each strata and obtained the corresponding y_{ij} values. The goal now is to select a second phase subsample from the elements of our first sample. This implies that respondents selected in the first SRSWOR are the only eligible candidates for selection in the second phase sample. Before drawing the phase 2 sample, we affiliate a Uniform(0,1) random number called uni with each interviewer in each strata. Then, within each strata, we sort the I interviewers by uni. This process results in a random permutation of interviewers within each strata.

For each interviewer, it is necessary to draw a simple random sample without replacement of size k or $k + 1$. Successful completion of this task involves creating another Uniform(0,1) random variable for all $n1h$ elements in the phase 1 sample. We label this random variable $unif$ and sort the data in two steps. For each strata, the data are sorted by interviewer, according to the random variable uni. Then, within each interviewer, the m elements are sorted by the variable $unif$. As a result, within each interviewer group of size m, there is a random ordering of the respondents. Finally, we choose a SRSWOR of size k or $k + 1$ from each interviewer.

Units for the phase 2 sample are selected from each of the I interviewers. Approximately the same number of units will be selected from each interviewer within a given strata. To be specific, define $n2h = k \cdot I + f$. If $f = 0$, then each interviewer randomly interviews k of the m available respondents. If, on the other hand, $f > 0$, the sampling process is more complicated. It is now necessary that we choose f of the I interviewers to have $k+1$ units randomly selected from each of their m units. Since the I interviewers are randomly ordered according to the variable *uni*, we assign the first f random interviewers to conduct $k+1$ interviews while the remaining $I - f$ interviewers conduct only k interviews. Again, we choose the *first* f interviewers simply for convenience.

Indi is an indicator variable for the I interviewers. If an interviewer is designated to conduct k interviews in the second phase sample, then $indi = 0$. However, if $f > 0$, then f randomly selected interviewers will sample $k+1$ units. These f interviewers have *indi* value equal to 1. Of course, the remaining $I - f$ interviewers have $indi = 0$.

The indicator variable *inds*2 is created to indicate membership in the second phase sample. Thus, *inds*2 equals 1 if that particular unit is an element of the phase 2 sample, and 0 otherwise.

The final step in this program is to output information on the variables of interest. For instance, all $N = 5,000$ population values of μ_r, x_r, and r are listed. From the phase 1 sample consisting of $n1h$ elements per strata, the variables y_{ij}, i, j, b_i, h, r, and *inds*1 are output. All elements of the phase 2 sample are automatically members of the phase 1 sample. The datasets *sample*2 and *sample*1 specifically list out the phase 2 and phase 1 sample elements, respectively.

References

[1] W. G. Cochran, *Sampling Techniques*, 3 ed., John Wiley & Sons, New York, 1977.

```
options ls = 80;
data one;

/* Define the parameters and set their values.  Note that some variable
   are created solely for the purpose of outputing the variables of
   interest in a specified order. */

seed = 454909;
cv = .7;
rhoy = .05;
rel = .8;
intbias = 0;
m1 = 1000;
respbias = m1*.10;
rhomux = .5;
mu0 = m1;
N = 25000;
I1= 20;
mm = 100;
n1 = I1*mm;
n2 = 1020;
I = I1;
m = mm;
alpha = ((1/cv**2)-1)/2;
exmu = 1000;
mumax = 2*exmu;
s2mu = (cv*exmu)**2;

bi = 0;
ii = 0;
j = 0;
mur = 0;
xr = 0;
yij = 0;

label cv = 'population coeff of variation';
label rhoy = 'interviewer correlation';
label rel = 'reliability';
label intbias = 'interviewer bias';
label respbias = 'respondent bias';
label rhomux = 'correlation between mu and x';
label mu0 = 'initial population mean';
label N = 'population size';
label n1 = 'phase 1 sample size';
label n2 = 'phase 2 sample size';
label I = 'number of interviewers';
label m = 'interviewer assignment size';
label yij = 'resp of ith interviewer, jth respondent';
label bi = 'interviewer effect';
label ii = 'ith interviewer: ii = 1 to I';
label j = 'jth respondent: j = 1 to m';
label mur = 'generated beta random var on 0 to 2000';
label xr = 'max of mur+er and 0';

  output;

  data two;set one;

* The variables of interest to the experimenter are kept and printed;

 keep cv rhoy rel intbias respbias rhomux mu0
      N n1 n2 I m;
proc print data = two;

/* We generate a beta(alpha,alpha) distribution on the support 0 to 2000
   To do this, we first generate a beta(alpha, alpha) distribution on th
```

```
   support 0 to 1 and call it zr.  Then multiplying by the constant, mum
   yields the appropriate distribution, call it mur.  The random variabl
   er is the respondent measurement error term. We repeat this process f
   all N values of the dummy variable r. We then generate the random
   variable er, with specified mean and variance.  Then, take the covari
   to be the max of 0 and mur+er.  We restrict the distribution of xr to
   be nonnegative since we require a nonnegative support.  */

data three; set one;
  array z(25000);
  array mu(25000);

do r = 1 to N;
p = ranuni(seed);
z(r) = betainv(p,alpha,alpha);
zr = z(r);
mu(r) = mumax*zr;
mur = mu(r);
er = rannor(seed)* sqrt(s2mu)*(1-rhomux)/rhomux;
xr = max(0,mur+er);

/* ur yields a random permutation of the population.  This permutation
   aids us when we take our phase 1 SRSWOR.  */

ur = ranuni(seed);
output;
keep  mur xr ur r n1;
end;
proc sort; by ur;

data four; set three;
if (_n_ > n1) then delete;

data five; set one;

/* We construct n1 yij values.  The mur values generated above correspon
   to the muij's in the model yij = muij + bi + epsij.  We now generate
   the interviewer and respondent error terms as follows. */

 do ii = 1 to I;
 bi = intbias + sqrt(rhoy*s2mu/rel)*rannor(seed);

   do j = 1 to m;
   epsij = respbias + sqrt((s2mu/rel)*(1-rel-rhoy))*rannor(seed);

 beij = bi + epsij;
 output;
end;
end;
keep beij ii j bi epsij;

/* We require the values of the Agriculture Survey response variable to
   be nonnegative.  Thus, we take yij to be the max of 0 and mur + beij

data six; merge four five;
yij = max(0,mur + beij);

/* We indicate which elements of our population are elements of the phas
   1 SRSWOR using the indicator variable inds1. */

data seven; set three;
if _n_ < n1+1 then inds1=1; else inds1=0;
keep mur xr inds1 r;

data eight; set six;
keep yij ii j bi;
```

```
data nine; merge seven eight;

/* We associate with each interviewer a random number, call it uni.  We
   then sort the n1 elements by the random number uni.  If l=0, then k
   units are selected at random from the m units of each of the I inter-
   viewers.  If l>0, then l of the I interviewers are selected at random
   first.  For these selected interviewers, k+1 units are selected at
   random from each of their m units.  For the remaining I-l interviewer
   k units are selected at random from each of their m units. */

data ten; set two;
seed = 237924;
do ii = 1 to I;
uni = ranuni(seed);
output;
end;
proc sort data = ten; by uni;

data eleven;
set ten;
k = int(n2/I);
l = n2-k*I;
indi = 0;

/* Indi is an indicator variable for the I interviewers.  If an
   interviewer has k units to be selected in the phase 2 sample, indi =
   For those l interviewers with k+1 units to be selected in the phase 2
   sample, indi has value 1. */

if _n_ < l +1 then indi = 1;
proc sort data = eleven; by ii;

data twelve; merge eight eleven; by ii;
data thirteen; set twelve;
seed = 3452801;

/* For each interviewer, we must draw a SRSWOR.  Therefore, we randomly
   order the m respondents within each interviewer group using the vari-
   able unif.  We then choose a SRSWOR of size k or k+1 from each inter-
   viewer, depending on the value of l. */

unif = ranuni(seed);
proc sort; by ii unif;

data fourteen; set thirteen;
inds2= 0;
if m*(ii-1) <  _n_ and _n_ < m*(ii-1) +k +indi +1 then inds2 = 1;
proc sort data=fourteen; by ii j;

/* This is the data set of N = 25,000 population values.  The indicated
   variables of interest and their values are retained. */

 data fifteen; merge seven fourteen;
keep mur xr yij ii j bi inds1 inds2 indi r;
proc sort ; by r;

/* The data sets sample1 and sample2 list out the variables of interest
   contained in the first and second phase samples, respectively.  The
   output data sets are given. */

data sample1; set fifteen;
if inds1=1;
proc sort; by ii j;

data sample2; set sample1;
```

```
if inds2=1;
proc sort; by ii j;

proc print data = sample2;
proc print data = sample1;

run;
```