

COMBINED RATIO ESTIMATES COMPUTED WITH SAS

Joan M. Pausche
United States Bureau of the Census

Two of the most important statistics provided by the Census Bureau for the Decennial Census are the mail response rate and the mail return rate. In general terms, the mail return rate is a measure of respondent participation and is defined as the number of census questionnaires returned to the Bureau by mail as a percentage of the number of occupied housing units. Note that each housing unit should have received one questionnaire. The mail response rate is the measure of housing units that are exempt from follow up, which includes vacant as well as occupied units. This rate is defined as the number of questionnaires returned by mail as a percentage of the number of occupied and vacant units that were mailed questionnaires.

Historically, these rates have been based on full census data. However, for a special study of 1980 census data, mail response and return rates were estimated from a one-stage, stratified, cluster sample of enumeration districts, which were units of census collection geography. Since each rate is a ratio of two variables which varied from sampling unit to sampling unit, a ratio estimator was used to obtain a 90 percent confidence interval for the estimate.

The national universe of enumeration districts, over 232,000, was separated into 11 geographic strata which were sampled independently using systematic random sampling. Each of the 11 strata contained 300 to 800 enumeration districts yielding a sample size of 6,793. Due to budget and time constraints, the proportion of stratum sample size to stratum universe was not the same for all strata. At the enumeration district level, data were gathered based on all the addresses in the enumeration district.

A combined ratio estimate, that is, a ratio of estimated stratum averages, was used to estimate the mail response and the mail return rates for the 1980 special study. Let y_i represent the number of questionnaire mail returns and let x_i represent the number of occupied housing units in a sampled enumeration district. Then the combined estimate of the mail return rate is obtained by computing the ratio

$$\frac{\sum_{h=1}^{11} N_h \bar{y}_h}{\sum_{h=1}^{11} N_h \bar{x}_h} \times 100$$

where N_h is the total number of enumeration districts in the h th stratum, $\bar{y}_h$ is the mean number of mail

returns in the h th stratum, and $\bar{x}_h$ is the mean

number of occupied housing units in the h th stratum. The mail response rate estimate can be obtained in the same way by letting x_i represent the number of occupied and vacant housing units. The estimate is multiplied by 100 since these rates are always reported as percentages.

The estimated variance of the mail return rate can be expressed in terms of the variance of d_{hi} , where

$$d_{hi} = y_{hi} - R x_{hi} \quad \text{and} \quad R = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

The variance of the combined mail return rate estimate, which is produced using the Taylor Series expansion method, is expressed as

$$\frac{1}{N^2 \bar{x}^2} \sum_{h=1}^{11} N_h^2 \frac{1-f_h}{n_h} s_{hd}^2, \text{ where}$$

$$s_{hd}^2 = \frac{\sum_{i=1}^{n_h} (d_{hi} - \bar{d}_h)^2}{n_h - 1}$$

$$= \frac{\sum_{i=1}^{n_h} \left[(y_{hi} - R x_{hi}) - \frac{\sum_{i=1}^{n_h} (y_{hi} - R x_{hi})}{n_h} \right]^2}{n_h - 1}$$

such that N is the total number of enumeration districts in the universe, n_h is the number of sampled enumeration districts in the h th stratum, and f_h is the sampling fraction (n_h/N_h) for the h th stratum.

SAS can be used to efficiently calculate the estimated variance or standard error (the square root of the variance) of the mail response and return rate estimates with basic procedures such as PROC SUMMARY. For standard errors of the response and return rates, SAS was used to find the following unknowns:

$$N^2, \overline{X}^2, N_h^2, f_h^2, n_h^2, S_{hd}^2. \quad \text{As seen later}$$

in a detailed description of the SAS code, N_h was passed in by using 11 macro variables, one for each stratum; N was calculated by summing these 11 variables. The remaining unknowns were calculated directly from the sample data or produced by SAS procedures.

First, the standard errors were calculated for the response and return rate estimates. The library name "joan" was defined for a VAX account using the following statement: `libname joan "[j_pausche]";`. The SAS code below names the macro, "jmrr," and identifies the 14 macro variables that were passed into the program. Since the response and return rate statistics were calculated for various subgroups of the population, all of the statistics were produced and printed within the macro so that the parameters could easily be changed to correspond to the subgroup data.

```
%macro jmrr(groupnm,subgrp,value,enc,esc,m,ma,
ne,p,sa,wnc,wsc,nyc,nyd);
```

In the macro definition of jmrr, the values of the variable groupnm include the names of the subgroups, including the overall population, and is used only for descriptive titles in the printed output. The variable subgrp identifies the subgroups (including the full population) and is used to separate subgroups from the sample data. The variable value is used in conjunction with subgrp to subset the sample data. The last 11 macro variables represent the values for N_h , the total number of enumeration districts in each stratum.

The SAS code below produces the temporary data set *fulldeck* from the sample data SAS file *joan.mrrsesug* and identifies y_i or $ytot$ for both the response and return rates, x_i or $xtot$ for the response rate, and x_i or $xotot$ for the return rate.

```
data fulldeck;
set joan.mrrsesug;
_type_=0;
if &subgrp=&value;
    ytot=s_mr+l_mr;
    xtot=s_occ+l_occ+s_vac+l_vac;
    xotot=s_occ+l_occ;
run;
```

Note that the value of `_type_` is "0" for every observation in the data set *fulldeck*. This serves two purposes which will be discussed later.

The subsetting IF statement in the 4th line creates the file *fulldeck* as a subgroup of *joan.mrrsesug* based on the values of the macro variables subgrp and value. To calculate statistics for the full population, "_type_" is passed into the code from the macro call as the value of the variable subgrp and "0" is passed through for the macro variable value to create the statement, `if _type_=0;`. This statement is true for every record in *joan.mrrsesug* due to line 3.

For both mail response and return rate standard errors, d_i was produced to obtain the variance of d_h . The SAS

code below calculates $\sum_{i=1}^n y_i$ and $\sum_{i=1}^n x_i$ using

PROC SUMMARY, which are then used to calculate R in the output data set *fullsum*.

```
proc summary data=fulldeck;
var ytot xtot xotot;
output out=fullsum sum=;
run;
data fullr(keep=_type_ rtot rotot);
set fullsum;
rtot=ytot/xtot;
rotot=ytot/xotot;
run;
```

Note that the overall sum of $ytot$ is also named $ytot$ by default since names were not specified in the SUM option of the output statement in line 3. PROC SUMMARY produced two automatic SAS variables, `_freq_` and `_type_`, which were included in the output data set *fullsum*. The values of the variable `_freq_` are the number of observations that were involved in producing each observation in the output data set and `_type_` contains information about the "class" variable. Since no class statement appeared in the SAS code, the value of `_type_` is "0." From lines 5-9, the variables

rtot and rotot are created using the sums produced by PROC SUMMARY. The variable rtot is R for the response rate and the variable rotot is R for the return rate. The data set *fullr* contained the variable *_type_* and the Rs for the two estimates.

The following SAS code creates the variate d_i where $d_i = y_i - Rx_i$, and computes the variance of d_i , which is

represented by S_{hd}^2 .

```
data full;
merge fulldeck fullr;
by _type_;
dtot=ytot-(rtot*xtot);
dotot=ytot-(rotot*xotot);
run;
proc sort data=full;
by stratum;
run;
proc summary data=full;
by stratum;
var dtot dotot;
output out=totdvar var=vtot votot;
run;
```

To create d_i or dtot for the response rate and dotot for the return rate standard error, the Rs (rtot and rotot) were joined with the data set *fulldeck* using the merge statement in line 2. Merging by the variable *_type_*, whose value is the same for every observation in *fulldeck* and *fullr*, assigns the same value for rtot for every observation in the data set *full*. Similarly, rotot has the same value for every observation. If the BY statement were omitted, the values of rtot and rotot would appear only in the first observation and would assign missing values to the remaining observations. Thus dtot and dotot representing d_i have a value for each sampled enumeration district.

In lines 7-9, the data set *full* was sorted by stratum to utilize PROC SUMMARY in lines 10-14 to compute variances of dtot and dotot for each stratum. The VAR option in the output statement in line 13 was used to produce the stratum variances of dtot and dotot and are named vtot and votot, respectively, in the output

data set *totdvar*, thus providing S_{hd}^2 .

The SAS code below introduces N_h by passing 11 macro variables for stratum population totals into the program and calculates the quantity to be summed over all strata.

```
data total;
set totdvar;
nnh=0;
if stratum="ENC" then nnh=&enc;
if stratum="ESC" then nnh=&esc;
if stratum="M" then nnh=&m;
if stratum="MA" then nnh=&ma;
if stratum="NE" then nnh=&ne;
if stratum="P" then nnh=&p;
if stratum="SA" then nnh=&sa;
if stratum="WNC" then nnh=&wnc;
if stratum="WSC" then nnh=&wsc;
if stratum="NYC" then nnh=&nyc;
if stratum="NYD" then nnh=&nyd;
fh=1-(freq/nnh);
itot=(fh*(nnh**2)/freq)*vtot;
iotot=(fh*(nnh**2)/freq)*votot;
run;
```

This data statement produces the data set *total* from *totdvar* and creates the variables nnh which represents N_h , fh which represents $1-f_h$, itot and iotot which

represent the quantity $N_h^2 \frac{(1-f_h)}{n_h} S_{hd}^2$. The

variable *freq* in lines 15-17 from the output summary data set *totdvar* equals n_h for the response and return rate standard errors; the data set *totdvar* contained one value of *freq* for each stratum.

The following SAS code sums itot and iotot over all strata and calculates the mail response and return rate standard errors.

```
proc summary data=total;
var itot iotot;
output out=totvar sum=ftot fotot; run;
proc summary data=fulldeck;
var xtot xotot;
output out=totmeans mean=mtot motot;
run;
data toterr;
merge totvar totmeans;
nn=&enc+&esc+&m+&ma+&ne+&p+&sa
+&wnc+&wsc+&nyc+&nyd;
stot=100*(sqrt(ftot/((nn**2)*(mtot**2))));
```

```

sotot=100*(sqrt(fotot/((nn**2)*(motot**2))));
label stot="Response Rate SE"
sotot="Return Rate SE";
keep _freq_ stot sotot;
run;

```

The output data set *totvar* created by the first PROC SUMMARY summed *itot* and *iotot* over all strata to

produce
$$\sum_{h=1}^{11} N_h^2 \frac{(1-f_h)}{n_h} S_{sd}^2$$
 for each standard

error. These are named *ftot* and *fotot* in the output summary data set *totvar*. The second PROC SUMMARY produced the means of *xtot* and *xotot* from the data set containing all of the sample data by specifying the MEAN option in the output statement. These were named *mtot* and *motot* in the output data

set *totmeans* and represent $\bar{X}$ for each standard error. The summary data sets were merged into the data set *toterr*, which contained one observation. In line 10 the variable *nn*, which represents the population total *N*, is introduced as the sum of the 11 stratum population totals. Finally, the response rate standard error (*stot*) and the return rate standard error (*sotot*) were calculated using the SAS square root function

"SQRT" and squaring each of N and $\bar{X}$ using the operator "***". The standard errors are multiplied by 100 to display the mail response and return rate standard errors as percentages.

Next, the mail response and return rate estimates were calculated using a data set produced during the standard error calculations. To produce the combined ratio estimates of the response and return rates, SAS was used to find the following unknowns:

$$N_h, \bar{Y}_h, \bar{X}_h.$$

The SAS code below produces the stratum means from the data set *fulldeck*, which are represented by *yntot*, *xntot*, and *xnotot*. These variables are then combined with *N_h* or *nnh* from the previously created data set *total* to produce *yjtot*, *xjtot*, and *xjotot*, which

represent $N_h \bar{Y}_h$ and $N_h \bar{X}_h$.

```

proc summary data=fulldeck;
by stratum;
var ytot xtot xotot;

```

```

output out=totnum mean=yntot xntot xnotot;
run;
data totmerg;
merge totnum total(keep=stratum nnh);
by stratum;
yjtot=nnh*yntot;
xjtot=nnh*xntot;
xjotot=nnh*xnotot;
run;

```

The stratum means were obtained by using the MEAN option in the output statement. The data sets *totnum* containing the stratum means and *total* were merged by stratum to create the new data set *totmerg*.

The following SAS code uses PROC SUMMARY to sum $N_h \bar{Y}_h$ and $N_h \bar{X}_h$ over the 11 strata and a data step to compute the mail response and return rate estimates as percentages.

```

proc summary data=totmerg;
var yjtot xjtot xjotot;
output out=totrat sum=yftot xftot xfotot;
run;
data totest;
set totrat;
etot=(yftot/xftot)*100;
eotot=(yftot/xfotot)*100;
label etot="Overall Response Rate Estimate"
eotot="Overall Return Rate Estimate";
keep etot eotot;
run;

```

Note that the SUM option in the output statement summed the product of the stratum means and the stratum population totals from the data set *totmerg* over all strata producing the variables *yftot*, *xftot*, and *xfotot* in the output data set *totrat*. The data set *totest*

was created from *totrat* where $\sum_{h=1}^{11} N_h \bar{Y}_h$ is divided

by $\sum_{h=1}^{11} N_h \bar{X}_h$ to produce the response rate estimate *etot* and the return rate estimate *eotot*.

The last step is to produce and print a 90 percent confidence interval containing each estimate. The 90 percent confidence interval for each estimate was

computed using the pivotal method; $r - (z \times SE_r)$

is the lower bound for the interval and

$r + (z \times SE_r)$ is the upper bound where r is the ratio estimate, $z=1.645$, and SE_r is the standard error of the estimate.

```
data totprn;
merge totest toterr;
z=1.645;
attrib utot uotot ltot lotot length=$4;
utot=left(round(etot+(z*stot),.1));
uotot=left(round(eotot+(z*sotot),.1));
ltot=left(round(etot-(z*stot),.1));
lotot=left(round(eotot-(z*sotot),.1));
attrib ctot cotot length=$11;
ctot=ltot || " - " || utot;
cotot=lotot || " - " || uotot;
label ctot="Response Rate CI"
cotot="Return Rate CI"
_freq_="Sample (n)";
run;
```

The data set containing the estimates *totest* was merged with *toterr*, which contained the standard errors, to produce the lower bounds for the confidence intervals, *ltot* and *lotot*, and the upper bounds, *utot* and *uotot*, as shown in lines 5-8. The SAS "ROUND" function was used to round the lower and upper bounds to one decimal place, the "LEFT" function was used to convert the bounds to character variables and left-justify the values, and the attrib statement was used to eliminate blank spaces. To produce print quality confidence intervals, the SAS character concatenation operator "||" was used to join the lower and upper bounds with a dash (surrounded by blank spaces) between them, as shown in lines 10 and 11.

The SAS code below printed 90 percent confidence intervals for the mail response and mail return rate with the appropriate title for each subgroup, as specified by the macro variable *groupnm* in line 4.

```
proc print data=totprn label noobs;
options nodate;
var _freq_ ctot cotot;
title "&groupnm Mail Response and Return Rate Estimates";
title2 "Using 90% Confidence Intervals (CIs)";
run;
```

The SAS statements below end the macro definition and call the macro three times to compute the estimates for various subgroups. The first macro call produces estimates for the entire population, whereas the second and third macro calls produce estimates for the "TAR" sub-population and the "Prelist" sub-population. Note that the last eleven variables, or the strata population parameters, reflect the population totals corresponding to each subgroup in the three macro calls.

```
%mend jmrr;

%jmrr(Overall,_type_,0,44216,15915,8102,19743,
11663,28698,40275,2002,25864,6375,11389)

%jmrr(TAR,tea,1,23768,4587,5491,11852,6892,
21617,17285,693124,6375,5804)

%jmrr(Prelist,tea,2,20448,11328,2611,7891,4771,
7081,22990,1327115440,0,5585)
```

The printed output of the SAS program is as shown below.

Overall Mail Response and Return Rate Estimates Using 90% Confidence Intervals (CIs)

Sample (n)	Response Rate CI	Return Rate CI
6793	73.3 - 74.1	82.3 - 82.9

TAR Mail Response and Return Rate Estimates Using 90% Confidence Intervals (CIs)

Sample (n)	Response Rate CI	Return Rate CI
3603	74.5 - 75.6	81.4 - 82.3

Prelist Mail Response and Return Rate Estimates Using 90% Confidence Intervals (CIs)

Sample (n)	Response Rate CI	Return Rate CI
3109	71.5 - 72.7	83.2 - 83.9

The output data shown were fabricated for demonstration purposes and are not the actual results of the study.

This paper reports general aspects of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

REFERENCES

Cochran, William G. Sampling Techniques. New York: John Wiley & Sons, 1977.

Hansen, Morris H., William N. Hurwitz and William G. Madow. Sample Survey Methods and Theory. New York: John Wiley & Sons, 1953.

SAS® User's Guide: Basics, Version 5 Edition, Cary, NC, 1985.

SAS is a registered trademark of SAS Institute, Inc., Cary, NC.