

The Use of SAS® to Construct Horvitz-Thompson Type Estimators for Safety Belt Compliance

Holmes Finch

Department of Statistics, University of South Carolina
Columbia, SC

ABSTRACT

In many survey applications, multi-stage sampling is used in an effort to reduce the variance and to obtain parameter estimates for very specific subgroups. The form of the estimators for such sampling schemes are complex and quite often not addressed by standard survey analysis software. One method for calculating these estimators has been derived and yields Horvitz-Thompson type estimates. It uses a recursive process which builds estimators at one stage on those from the stages below, employing a combination of weights and raw estimates from individual sites. These estimators were used to calculate the proportion of vehicle occupants wearing safety belts in South Carolina during the Fall of 1996, using data gathered from a 3 stage sample observational survey. SAS was employed to calculate both the estimate of compliance and the standard error of the estimate.

INTRODUCTION

As a part of a program initiated by President George Bush in 1990, the state of South Carolina (SC) has sponsored a state-wide observational study of safety belt use each year from 1991 through 1996. The Statistical Laboratory (StatLab) at the University of South Carolina has conducted this survey each of these years, producing an estimate for the state which is then forwarded to officials at the U.S. Department of Transportation. In 1990, the state legislature passed and the governor signed legislation requiring that all front seat occupants of motor vehicles wear a safety belt, with non-compliance punishable by a fine. However, this was passed as a secondary law, meaning that individuals could not be pulled over for not wearing their safety belt, but those not wearing them when they were pulled over for another offense could be charged

with failure to wear the safety belt as well.

In order to determine the percentage of the state's residents using safety belts for reporting to the Federal Government, and to assess the effectiveness of an advertising campaign encouraging their use, the Department of Public Safety (DPS) has sponsored the state-wide survey of safety belt use. The survey, originally conducted by a marketing research company, was first done by the StatLab in May, 1991. The original methodology matched that used by the marketing firm in an attempt to make the new results comparable with the old. However, in 1993 the U.S. Department of Transportation initiated new guidelines which were to be used by all agencies conducting such surveys when the results were to be used as official documentation in conjunction with the Bush initiative. These guidelines fundamentally altered the methods used by the StatLab in the conduct of the survey, particularly in the area of sampling. The new sampling plan was much more complex than the original one, and required the development of completely new estimators of safety belt use. These estimators, which were not available in standard survey analysis software, are based on calculations by Horvitz and Thompson and were derived by staff at the StatLab. It was determined that SAS would be used to encode the calculation of these estimators and their variances. In this paper, we would like to address the sampling plan used to select sites for the survey, the mathematical form of the estimators and their variances, and the SAS code used to obtain the estimates.

METHODOLOGY

The methodology described below represents what we consider to be the best sampling scheme available, while

maintaining standards established by the National Highway Traffic Safety Administration (NHTSA) in 1993. Using stratification and multi-stage probability sampling, 16 of South Carolina's 46 counties were selected for inclusion in the sample. The number of counties selected for observation was based upon recommendations by the National Highway Traffic Safety Administration (NHTSA) (Federal Register, 1992).

The first layer of stratification was at the state level, with the initial strata being geographic and demographic in nature. The state was divided into 3 geographic regions (the coastal counties, the midlands counties, and the up-state counties) which were roughly equal with respect to average traffic intensity as provided by the SC Department of Transportation (DOT). The counties were further divided by whether the majority of their populations were classified as living in urban or rural areas. Thus, there were a total of 6 strata statewide from which sampling could occur. Based upon the distribution of average vehicle miles traveled, the number of counties to be chosen from each strata was as follows: 4 coastal urban, 4 midlands urban, 3 upper urban, 2 coastal rural, 2 midlands rural, 1 upper rural. Within these strata, individual counties were selected randomly with probability of selection proportional to average vehicle miles traveled.

Within each selected county, 4 census tracts were randomly sampled with probability proportional to average vehicle miles traveled. The number of tracts selected was largely a function of logistics; the budget constrained the amount of time that could be spent in a county (1 day), and the amount of daylight constrained how long such a day could be. Finally, within census tracts, 3 individual observation points were selected. The DOT positions electronic traffic counters along roads throughout S.C., each of which yields average traffic intensity data for a 24 hour period. These

numbers were combined to obtain the average intensities for strata, counties and census tracts, and were used to weight individual sites for sampling within the tracts. The actual point of observation was placed as close to the counting site as possible.

The derivation of the estimators will not be discussed here, however we will examine their form at each level of calculation. At the state level, the estimate of the proportion of vehicle occupants wearing safety belts takes the standard form for stratified samples:

$$P_{sc} = \sum_{h=1}^H W_h p_h$$

where

P_{sc} = Proportion of occupants wearing safety belts in SC

W_h = Average traffic intensity for strata h

p_h = Proportion of sampled occupants wearing safety belts in strata h

Within strata, counties and census tracts, the estimators can be expressed as:

$$p_h = \frac{\sum_{c=1}^C \frac{a_c}{w_c}}{\sum_{c=1}^C \frac{m_c}{w_c}}$$

where

a_c = Number of occupants observed wearing safety belts in sampling unit c

m_c = Number of occupants observed in sampling unit c

w_c = Weight for sampling unit c

The estimators for the sites are calculated in the expected fashion.

$$p_i = \frac{a_i}{m_i}$$

where

a_i = Number of occupants wearing
safety belts at site i

m_i = Number of occupants observed
at site i

While the form of the estimators themselves are not very complex, the calculation of the appropriate weights does present some computational challenges. These weights indicate the likelihood that a particular sampling unit (site/tract/county) was selected from all of the possibilities in the sampling frame. Therefore, it is necessary to have the respective traffic intensities for each of the *possible* sampling units. Below are the equations for the calculation of these weights when 1, 2, 3 and 4 units are selected from the pool of possible units.

In addition to estimating the proportion of vehicle occupants wearing safety belts, it is also necessary to estimate the standard error of the statewide proportion. The agency sponsoring the research was not interested in the standard errors for any estimators other than the state, which necessitated the calculation of standard errors for each of the strata in order to obtain the statewide estimate. At the end of this paper, the formula for the overall variance is displayed in its entirety.

CONCLUSION

The calculation of estimators in complex, multi-stage sample surveys can be very problematic. Generally, standard approaches do not take into account the complexity of the sampling design, particularly when the methods used to sample change across the stages. In our experience, standard computer packages allow for perhaps 2 levels of stratification, but we have not found any software that allows for 3 levels as we have in the safety belt survey. The derivation of the estimators presented here allows for a calculation of safety belt

use that is hopefully much closer to reality than one that is based on a simpler sampling scheme. By using SAS macro capabilities, much of the computer work has become automated, allowing the relatively quick turnaround from the point of data entry to the calculation of the estimates of interest.

Further work is still to be done on the calculation of standard errors, in an attempt to account for the variability in the number of vehicles actually included in the survey. An implicit assumption in the current calculations is that at the site level, all vehicles will be counted, which is true at some intersections. However, at most intersections the traffic is too busy for such a complete survey, and only a portion of all possible vehicles are counted by the observers. Ideally, the estimate of variance would allow for some flexibility in this regard, and it is hoped that adjustments can be made to correct for this factor in future iterations of the study.

SAS CODE

Below is included a sample of the SAS code used to calculate the estimators described in this paper. The complete programs are available from the author upon request.

Strata Estimates

```
%macro stratapt(strata);
data all1;
    infile 'c:\data\county.dat' delimiter=' ' missover;
input county $ strata $ tract site gender $ place $
buckled $ count;
if strata="&strata";

libname total 'c:\data';
data wt1;
    set total.sitewts;      **READS IN SITE
WEIGHTS**;
    *input county $ tract site weight;
    if strata="&strata";
proc means noprint;      **COMPUTES THE
TOTAL NUMBER BUCKLED BY SITE**;
var count;
output out=sumbelt1 sum=belted;
by site;
```

```

id tract;
proc sort data=all1;
  by site;
proc means noprint;  **COMPUTES THE
TOTAL NUMBER COUNTED BY SITE**;
```

var count;

output out=sumoccl sum=occupant;

by site;

id tract;

data buckwt1; **COMPUTES

THE WEIGHTED SITE TOTALS;

merge wt1 sumbelt1 sumoccl;

by site;

bHat = belted/weight;

tHat = occupant/weight;

The weighted totals for the tracts and counties are calculated in similar fashion.

```

data strata;
  merge countpi sumtrac2;  **COMPUTES
WEIGHTED COUNTY TOTALS**;
```

by county;

sbHat = bbHatsum/weight2;

stHat = ttHatsum/weight2;

proc sort; by county;

proc means noprint; **SUMS WEIGHTED

COUNTY TOTALS ACROSS STRATA**;

var sbHat stHat;

output out=stratest sum = sbHatsum

stHatsum;

data final;

set stratest; *****COMPUTES

STRATA PERCENTAGE*****;

strpct = sbHatsum/stHatsum;

Calculation of W_{ij}

****CALCULATE ZJ TERMS**;**

```

do i=1 to 3;
  sum1j=0; sum2j=0;
  ind=i;
  ptempi=p{ind};
do j=1 to &n;
  if j ne i then do;
    ind=j;
    ptempj=p{ind};
    zj=ptempj/(1-ptempj);
    sum1j=sum1j+zj;
    sum2k=0;
```

****CALCULATE ZK TERMS**;**

```

do k=1 to &n;
  if k ne j and k ne i then do;
    ind=k;
    ptempk=p{ind};
```

```

    zk=ptempk/((1-ptempj-ptempk)*(1-
ptempk));
    sum2k=sum2k+zk;
end;
end;
sum2j=sum2j+ptempj*sum2k;
end;
end;
final1=ptempi*(1+sum1j+sum2j);
output;
end;
```

Calculation of W_{ij}

****CALCULATE ZJ TERMS**;** **Begins

computing the PIE-ij s**;

```

do i=1 to 3;
  ind=i;
  zi=p{ind};
  tempi=1-zi;
do j=1 to 3;
  if j gt i then do;
    ind=j;
    zj=p{ind};
    tempj=1-zj; tempij=1-zi-zj;
    k0=0; k1=0; k2=0; k3=0; k4=0;
    p1=0; p2=0; p3=0;
**CALCULATE ZK TERMS**;
do k=1 to &n;
  if k ne j and k ne i then do;
    ind=k;
    zk=p{ind};
    tempk=1-zk; tempik=1-zi-zk; tempjk=1-zj-
zk;
    k0=k0+zk; k1=k1+zk/(tempk*tempik);
    k2=k2+zk/(tempk*tempjk);
    k3=k3+zk/tempik; k4=k4+zk/tempjk;
  end; end;
p1=(zi*zj/(tempij))*(k0*((1/tempi)+(1/tempj)));
p2=(zi*zj)*(k1+k2); p3=(zi*zj)*(k3/tempi
+ k4/tempj);
final1=p1+p2+p3;
output;
end; end; end;
```

REFERENCES

Cochran, William G. (1977) *Sampling Techniques*, New York: John Wiley and Sons.

Thompson, Steven K. (1992) *Sampling*, New York: John Wiley and Sons.

RELEVANT FORMULAS

Weights

The weights used in the calculation of the estimates are calculated based upon the number of units selected. Below are the 4 possible manifestations of these weights. Following are the formulas used in the calculation of the variance at the strata level. Note that the calculation of statewide estimates was done using the standard equations on Excel®.

One unit

$$w_{i1} = \frac{z_i}{\sum_{i=1} z_i}$$

where

z_i = Traffic intensity for unit i

Two units

$$w_{i2} = w_{i1} \left(1 + \sum_{j \neq i} \frac{w_j}{1 - w_j} \right)$$

where

w_j = Proportion of traffic intensity for unit $j \neq i$ (calculated as was w_{i1})

Three units

$$w_{i3} = w_{i1} \left(1 + \sum_{j \neq i} \frac{w_j}{1 - w_j} + \sum_{j \neq i} \left(w_j \sum_{k \neq i, j} \frac{w_k}{(1 - w_k)(1 - w_j - w_k)} \right) \right)$$

Four units

$$w_{i4} = w_{i1} \left(1 + \sum_{j \neq i} \frac{w_j}{1 - w_j} + \sum_{j \neq i} \left(w_j \sum_{k \neq i, j} \frac{w_k}{(1 - w_k)(1 - w_j - w_k)} \right) + \sum_{j \neq i} w_j \left(\sum_{k \neq j, i} w_k \sum_{l \neq k, j, i} \frac{w_l}{(1 - w_k)(1 - w_j - w_k)(1 - w_j - w_k - w_l)} \right) \right)$$

$$\text{Variance} = \frac{1}{x^2} \left(s + \frac{b}{I} \right)$$

where

$$s = \sum_{j=1}^I \sum_{i>1} \frac{(w_i w_j - w_{ij}) \left(\frac{d_i}{w_i} - \frac{d_j}{w_j} \right)^2}{w_{ij}}$$

$$d_i = a_i - pm_i$$

$$w_{ij} = g_1 + g_2 + g_3$$

$$x = \sum \frac{m_i}{w_i}$$

$$g_1 = \left(p_i \left(\frac{p_i}{1 - p_i - p_j} \right) \right) \left(p_k \left(\left(\frac{1}{1 - p_i} \right) + \left(\frac{1}{1 - p_j} \right) \right) \right)$$

$$g_2 = (p_i p_j) \left(\sum \left(\frac{p_k}{(1 - p_k)(1 - p_i - p_k)} \right) + \left(\frac{p_k}{(1 - p_j - p_k)} \right) \right)$$

$$g_3 = (p_i p_j) \left(\left(\frac{\sum \frac{p_k}{(1 - p_i - p_k)}}{(1 - p_i)} \right) + \left(\frac{\sum \frac{p_k}{(1 - p_j - p_k)}}{(1 - p_j)} \right) \right)$$

$$x = \sum_{i=1}^I \frac{m_i}{p_i}$$

$$b = \sum_{i=1}^I f^2 \left(1 - \frac{4}{f} \right) \left(\frac{\text{var}_i}{p_i} \right)$$

f = number of sub - units in sampling unit i

var_i = simple variance of sampling unit i , pXq

Author's address

Holmes Finch
Department of Statistics
University of South Carolina
Columbia, SC 29208
(803)777-5074
finch@stat.sc.edu

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

Excel® is a registered trademark of
Microsoft, Inc. in the USA and other
countries.