Lab5 Cluster Sampling

Introduction

In previous labs, we dealt with sampling problems where the observation units were directly sampled. In this lab section, at first, we explore problems where groups of observation units, called clusters, are selected and units within selected clusters are observed.

Often though, there is little gain from sampling every unit within a selected cluster due to correlations within a cluster, so we may be better off selecting more clusters and fewer units within selected clusters. Also, when travel costs are negligible, it may be too expensive to observe all units within a cluster relative to the
Goals:

1) To become familiar with cluster sampling terminology.
2) To be able to perform one- and two- stage cluster estimation.

Example 1: GPA study

```
/* Analyzes the data in Example 5.2 of
   Sampling: Design and Analysis, 2nd ed. by S. Lohr
   Copyright 2008 by Sharon Lohr */
/* one-stage cluster sampling: equal cluster size*/
options ls=78 nodate nocenter;

data gpa;
   input suite gpa;
   wt = 20;  /* every person has weight 100/5 = 20 */
   datalines;
1 3.08
1 2.60
1 3.44
1 3.04
2 2.36
2 3.04
2 3.28
2 2.68
3 2.00
3 2.56
3 2.52
3 1.88
4 3.00
4 2.88
4 3.44
4 3.64
5 2.68
5 1.92
5 3.28
5 3.20
;
proc print data=gpa;
run;


proc boxplot data=gpa;
```

```
   plot gpa*suite;
run;


 /* To analyze a cluster sample, need statements for cluster and weight.*/
 /* Note that total (for the fpc) is to specify the number of psu's, not
     the number of observation units */
 /* As always, need weight variable to get correct answer for sum */

proc surveymeans data=gpa total = 100 nobs mean sum clm clsum;
   cluster suite;
   var gpa;
   weight wt;
run;


/* What do we get in an erroneous analysis that ignores the clustering?*/

 /* DO NOT USE THE FOLLOWING 2 LINES; THEY ARE INCLUDED TO SHOW YOU WHAT
HAPPENS
    WITH AN INCORRECT ANALYSIS */

proc surveymeans data=gpa nobs mean sum clm clsum;    /* This is wrong since
it does not include cluster or weight statements */
   var gpa;
   run;
 /* This analysis assumes that we have a SRS!  WRONG WRONG WRONG! */
```

Example 2: Algebra classes:

```
/* Analyzes the data in Example 5.6 of Sampling: Design and Analysis, 2nd ed.
   by S. Lohr. Copyright 2008 by Sharon Lohr */
/* One-stage cluster sampling: unequal cluster size */

filename algebra 'D:\TEACHING\T_STAT574\lab\lab4\algebra.csv';

options ls=78 nocenter nodate;
data algebra;
   infile algebra delimiter= ',' firstobs = 2;
   input class Mi score;
   sampwt = 187/12;
run;

proc print data=algebra;
run;

proc surveymeans data=algebra total = 187 nobs mean sum clm clsum df;
   cluster class;
   var score;
   weight sampwt;
   ods output Statistics=myout;
run;

proc print data=myout;
run;
```

Example 3: coots study

```
/* Analyzes the data in Example 5.7 of Sampling: Design and Analysis, 2nd ed.
   by S. Lohr
   Copyright 2008 by Sharon Lohr */
/*two-stage cluster sampling*/

filename cootsdat 'D:TEACHING\T_STAT574\lab\lab4\coots.csv';

options ls=78 nodate nocenter;
data coots;
   infile cootsdat delimiter=',' firstobs=2;
   input clutch csize length breadth volume tmt;
   relwt = csize/2;
   proc print data=coots;
   run;


 /* Construct the plot in Figure 5.3.  You can omit the goptions statements
    and the options in the plot statement if you wish to use the SAS plot
defaults. */

goptions reset=all;
goptions colors = (black);
symbol1 value= dot h = .5;
axis4 label=(angle=90 'Egg Volume') order=(0 to 5 by 1);
axis3 label=('Clutch Number') order =  (0 to 200 by 50);

proc gplot data=coots;
   plot volume * clutch/ haxis = axis3 vaxis = axis4;
run;

 /*  For the coots data we do not know the number of units in population */
 /*  Must include the weight variable since observation units have
     unequal weights. Here we use the relative weights */

proc surveymeans data=coots;
   cluster clutch;
   var volume;
   weight relwt;
run;
```