SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Using SURVEYSELECT to Draw Stratified Cluster Samples with Unequally Sized Clusters

#SASGF

Using SURVEYSELECT to Draw Stratified Cluster Samples with Unequally Sized Clusters

Chuanjie "George" Liao; Brian F. Patterson

Pearson, Iowa City, IA; Questar Assessment, Inc., Minneapolis, MN

ABSTRACT

- PROC SURVEYSELECT is a useful procedure for sample selection for a wide variety of applications. This paper presents the application of PROC SURVEYSELECT to a complex scenario involving a state-wide educational testing program, namely drawing inter-dependent stratified cluster samples of schools for the field-testing of test questions, which we call "items". These stand-alone field tests are given to only small portions of the testing population and as such, a stratified procedure was used to ensure representativeness of the field-test samples. As these items' field test statistics are evaluated for use in future operational tests, an efficient procedure is needed to sample schools, while satisfying pre-defined sampling criteria and targets. This paper provides an adaptive sampling application and then generalizes the methodology as much as possible for potential use in more industries.

METHODS

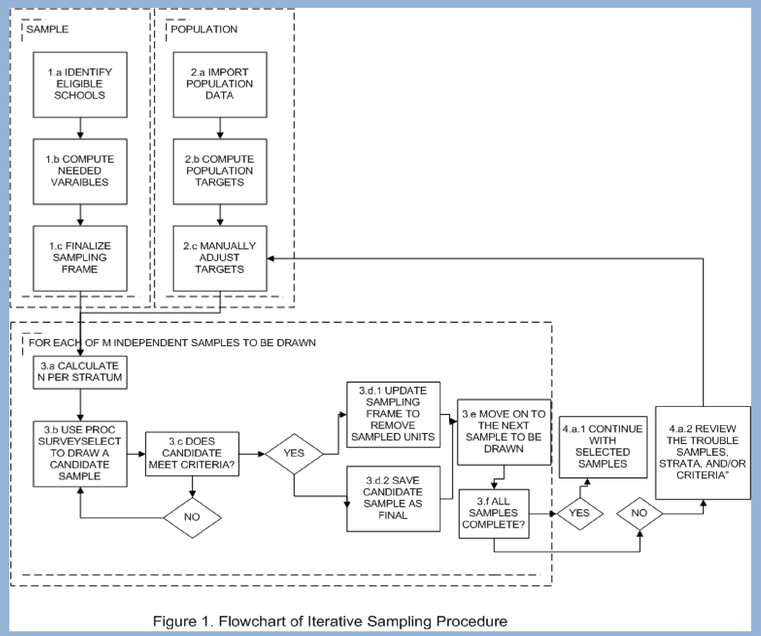

Figure 1. Flowchart of Iterative Sampling Procedure

METHODS CONTINUED

The first step is to prepare the sampling frame. From all schools in one state, only eligible schools can be selected as a sampling frame. The general rule for sampling in our example is to draw schools so that the whole drawn samples roughly match the prior year's population test score means for the relevant grade and subject, i.e. the entire state operational data population.

We create a strata variable School_Geog_SS_Qnt at school level, which is the combination of School_Geog and the quantile for the school's prior scale score means. School_Geog values are obtained from the sampling frame but scale scores are from prior year's population-wide test data.

In our example, we specify a dataset containing sampling targets (i.e. allocation variable) with our allocation weights via the ALLOC option to allocate the total sample across the strata. We need to assign related sampling weights after the corresponding strata variable is constructed and also adjust the sampling weights using the mean expected test-takers by subject, grade, and school geographic/urbanicity variable.

```
PROC SURVEYSELECT NOPRINT DATA= Frame_&Sample_Num._&Subject._&Gr._&Form
                  METHOD= SRS
                  N= &n_Schools
                  OUT= Sample_&Sample;
        STRATA School_Geog_SS_Qnt /
                  ALLOC= Target_OP_Year_Alloc_&Subject._&Gr._&Form;
        ID Subject Grade Form_Type School_ID School_Geog;
RUN;
```

Using SURVEYSELECT to Draw Stratified Cluster Samples with Unequally Sized Clusters

Chuanjie "George" Liao; Brian F. Patterson

Pearson, Iowa City, IA; Questar Assessment, Inc., Minneapolis, MN

METHODS CONTINUED

The sampling data frame Frame_&Sample_Num._&Subject._&Gr._&Form is classified by subject, grade and form in the ith drawn sample, and we have two forms usually – multiple-choice form and open-ended question form. So the allocation datasets Target_OP_Year_Alloc_&Subject_&Gr._&Form. are children datasets from the Target_OP_Year_Alloc dataset, by Subject, Grade and Form.

n_Schools indicates the total number of primary sampling units (PSUs). In our case, it refers to school-grade combinations.

SRS (simple random sampling) is used.

We evaluate the drawn samples from three criteria – total number of students, grand mean of scale scores, and school geographic targets. We have a target number of sampled examinees and a target absolute deviation proportion. In our practice, we pre-set the target scale score mean tolerance as 3 points. For each School_Geog category, we calculate the absolute School_Geog proportion deviation. If the largest absolute deviation is within the pre-determined 5% range, it is acceptable.

The whole sample drawing process was implemented in a repetitive loop and did not stop until either the sampling criteria were all met or we hit a pre-defined number of maximum iterations. If the sample passes the evaluation procedure, we update the sampling frame to remove those selected PSUs.

RESULTS

For the sample drawing purpose, we were able to balance computation time, precision of meeting sampling criteria, and other factors. When constructing strata variables, in our case, we only stratified schools into 5 quantiles (i.e., using 20th, 40th, etc. percentile ranks) and that number can vary by setting. For the sake of convenience, we then concatenated the scale score mean quantile with school geographic/urbanicity information to construct a single strata variable. This helped simplify the manner in which we assigned sampling weights while still achieving the sampling representativeness requirement. During the iterations of whole sample drawing, we do not just rely on PROC SURVEYSELECT as the main option, but also make adjustment from three perspectives - adjust number of clusters, adjust sampling weight and adjust sampling order to aid the drawing.

CONCLUSIONS

We have presented the application of PROC SURVEYSELECT in an educational testing program to select sample schools based on a combination of different sources of information and incorporated the general methodology and hands-on adjustment to meet strict state testing requirements. From demonstrating the flexibility and applicability of the procedure, we have provided some guidance for readers who would like to draw a cluster sample with unequally sized clusters with varying sample weights in any field.

REFERENCES

Lewis, T. (2013). "PROC SURVEYSELECT as a Tool for Drawing Random Samples," Paper presented at the annual conference of the Midwest SAS Users Group. Columbus, OH, September 22-24. Available online at: http://www.mwsug.org/proceedings/2013/AA/MWSUG-2013-AA02.pdf.

Lohr, S. (1999). Sampling: Design and Analysis. Pacific Grove, CA: Duxbury Press.

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF