Discussion: Applications of Sampling

Michael P. Cohen

Senior Consultant to NORC
1801 Pennsylvania Avenue NW
Washington DC 20006-3608
MPCohen@Juno.com

Introduction

The three presentations in this session illustrate the diversity in the ways sampling can be applied. All three of them introduce interesting new ideas rather than just refining conventional approaches.

I shall simply discuss the papers one by one in the order presented.

"Evolutionary Algorithms for Optimal Sample Design" by Charles Day

This paper employs an algorithm, called the cooperative coevolutionary algorithm (CCEA), with a strong intuitive appeal. Let me first summarize Day's CCEA in terms of what needs to be given in advance and what is produced.

Given in advance:
Measure of size (e.g. Adjusted Gross Income)
Number of strata *H*
Simple random sampling within a stratum
Maximum Coefficient of Variation (CV) for each of several pre-specified variables
Produced by algorithm:
Stratum boundaries
Allocations (stratum sample sizes)

For comparison purposes, we now give the corresponding information for the well-known algorithm of Lavallée and Hidiroglou (1988).

Given in advance:
Measure of size
Number of strata *H*
Simple random sampling within a stratum
Maximum CV for a *single* pre-specified variable
Produced by algorithm:
Stratum boundaries
Allocations (stratum sample sizes)

Let's also compare with the recent and quite different method of Gunning and Horgan (2004).

Given in advance:
Measure of size and number of strata *H*
Simple random sampling within a stratum
Maximum CV for a *single* pre-specified variable
Simple method but positive skew in population *required*
Produced by algorithm:

Stratum boundaries
Allocations (stratum sample sizes)

The Gunning and Horgan algorithm is unusual in actually requiring positive skewness in order to function correctly. But when there is positive skewness, it is particularly simple relative to other methods.

In comparing CCEA to the others, one sees that it is the only one that can control the CVs for several pre-specified variables at a time, a clear advantage. The intuitive appeal of the heuristic algorithm is another advantage.

I will close this section with two questions, possibly entailing additional research:
(1) Can number of strata H be allowed to vary (subject to a penalty)?
(2) Can we allow probability proportional to size (pps) sampling within strata?

"Derivation of Sample Size Formula for Cluster Randomized Trials with Binary Responses Using a General Continuity Correction Factor and Identification of Optimal Settings for Small Event Rates" by Majnu John and Madhu Mazumdar

This paper treats cluster randomized trials (CRTs). I was struck by the similarity of the issues and techniques with those used in survey sampling. In some cases the terminology is slightly different, but not difficult for a survey sampler to figure out. The unifying theme is the intraclass correlation that arises from clustering.

The aim of this research is determine the sample size, by which is meant the number of clusters, needed to guarantee no more than specified type I and II errors in doing a hypothesis test in doing a hypothesis test of the difference of two proportions. The authors develop a new formula for determining the sample size, (assuming equal sample size per cluster, an assumption made for tractability) and demonstrate its effectiveness. The example applications are very interesting in their own right.

We summarize the method in terms of what needs to be given in advance and what is produced:

Given in advance:
Binary outcomes
Population means, type I error (alpha), power (one minus type II error)
Design effects (variance inflation factors, VIFs)
Continuity correction factor c
Sample size per cluster (same for each cluster)
Produced by program:
Sample size (number of clusters) K

The program is implemented in R code that the authors supply. I tried it and it works. Supplying the code makes it much more likely that practitioners will use it.

Here are two questions to close the section:
(1) Why not use the "exact" method? (The authors, as I understand it, are working on this.)
(2) Can the equal-sized cluster assumption be relaxed?

"Sampling from Discrete Distributions: Application to an Editing Problem" by Lawrence Cox and Marco Better

This paper studies 2 by 2 contingency tables with count data missing for some cells but with marginal totals. It studies this problem with an eye toward more complex tables. We summarize the situation thusly:

Given in advance:
Counts of responses by category and nonresponses to two questions
Produced by algorithm:
Allocations of the non-responses that satisfy marginal constraints
Method that "scales up" to higher dimensions

Raking is one method that one might think of to tackle this problem. As the authors note, "…ratio adjustment and raking are ineffective, as they create imputed cell counts less than observed partial counts." Cox and Better develop version of raking that fixes this shortcoming

Obtaining a sample of the solution-feasible space is goal for future research. If probabilities of selection of each "point" in the solution-feasible space were known, one could calculate the variance due to imputation, confidence intervals, and related quantities.

Conclusion

This session has provided three very different applications of sampling, yet each is very innovative in its own way.

Acknowledgements: I sincerely thank the authors/presenters not only for their excellent presentations but also for their well prepared papers given to me in advance. Very special thanks to Dale Atkinson for organizing the session, chairing it, and inviting me to discuss.

References

Pierre Lavallée and Michael A. Hidiroglou, On the Stratification of Skewed Populations, *Survey Methodology*, 1988, Vol. 14, No. 1, 33-43

Patricia Gunning and Jane M. Horgan, A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations, *Survey Methodology*, 2004, Vol. 30 , No. 2, 159-166