Survey Errors and Survey Costs: Experience from Surveys of Arrestees

Y. Michael Yang

NORC at the University of Chicago, 1350 Connecticut Ave., N.W., Washington, DC 20036

KEY WORDS: Sample Allocation, Error Model, Cost Model

I. Introduction

Sponsored by the National Institute of Justice (NIJ), the Arrestee Drug Abuse Monitoring program (ADAM) conducted data collection in 43 urban counties (sites) and more than 100 booking facilities across the United States between 2000 and 2003. With the introduction of probability sampling, ADAM represented a significant improvement over its predecessor, the Drug Use Forecasting (DUF) system that used convenient samples. Within each site, the sampling of adult booking facilities and male arrestees within these facilities were carried out based on probability principles to the extent possible.[1] However, important aspects of the ADAM design still represented compromises with real world constraints. In particular, ADAM sample allocation was to a large extent dictated by the cost structure. Efficient sample allocation among facilities requires the specification of error and cost models that describe the errors and costs as a function of the sample design. In principle, the proportional allocation method was used to allocate the ADAM sample among sites, among facilities within sites, and among strata within facilities. In reality, the actual ADAM sample allocation, especially allocation among facilities, could deviate significantly from proportional allocation due to cost considerations. This paper discusses the error and cost models that determined the actual sample allocation among facilities.

Section II discusses the error and cost models under single-stage designs where each facility is a stratum. Section III presents the error and cost models under two-stage designs where each facility is a cluster (PSU). Section IV provides some concluding remarks.

[1] ADAM used convenience samples for female arrestees in adult facilities and for all detainees in juvenile facilities. Therefore, the following sample design and weighting discussions apply to adult male arrestees only.

II. Errors and Costs Under Single-Stage Designs

ADAM sample designs were site-specific because the structure of the arrestee population, among other things, varied by site. The most important determining factor was the number of booking facilities within a site. A single-stage design was used in sites where the number of facilities was small so that all the facilities could be included in the sample. When a site had a single booking facility, facility sampling was obviously irrelevant and all resources of the site were devoted to sampling arrestees from that one facility. When a site had two to four facilities, each facility was considered a sampling stratum, and the sampling of arrestees was carried out independently within each facility. In both cases, all facilities within a site were included in the sample, and arrestees were sampled in one stage.

The total site sample size was determined by the site budget. We consider the sample allocation among $m (1 < m \leq 4)$ facilities of a site where each of the m facilities is a stratum. For a variable Y, the sampling variance of the sample mean is

$$V_{st}(\hat{\bar{Y}}) = \sum_{h=1}^{m} \frac{(1-f_h) W_h^2 S_h^2}{n_h} \quad [1]$$

where

f_h = sampling fraction used in stratum h;
W_h = proportion of the frame population in stratum h;
S_h^2 = population element variance of stratum h;
n_h = sample size of stratum h.

For stratified sampling, a typical cost model may be expressed as

$$C_{st} = C_0 + C_1 m + \sum_{h=1}^{m} C_{2h} n_h \qquad [2]$$

where

C_0 = total fixed costs per site;

C_1 = additional fixed costs per facility;

C_{2h} = cost of sampling and interviewing a single arrestee from stratum h.

The fixed costs per facility may vary slightly, but that will not change the general structure of the cost model since the first two terms on the right hand side can be combined to represent the fixed costs per site.

If this cost model is accurate, the sampling error of an estimated population mean is minimized when the sample size per stratum n_h is made proportional to

$$\frac{W_h S_h}{\sqrt{C_{2h}}} \qquad [3]$$

Expression [3] represents the optimal allocation under the specified error and cost models (Cochran, 1977).

However, the ADAM interviewing costs were determined by the number of interviewers per facility rather than the number of interviews completed. ADAM collected data through face to face interviews at each sample facility. Each sample facility hired at least one interviewer who was paid 8 hours per day regardless of how many interviews were completed. Therefore, the following cost model is more appropriate

$$C'_{st} = C_0 + C_1 m + \sum_{h=1}^{H} C'_{2h} n'_h \qquad [4]$$

where

C'_{2h} = cost per interviewer in stratum h;

n'_h = number of interviewer employed in stratum h.

The number of complete interviews per facility is a discrete function of the number of interviewers and the size of the facility. An interviewer could conduct about 12 interviews per 8-hour shift. ADAM required that the data collection period in all facilities cover at least 1 week so that every weekday would be represented by the sample. Since every sample facility used at least one interviewer regardless of the size of the facility, the minimum allocation to each facility was 84 interviews. Within each facility, the relationship between the number of interviews and the number of interviewers may be expressed as follows,

$$n_h = \begin{cases} n'_h * w * 84 \\ N_h \end{cases} \qquad [5]$$

where

w = number of weeks of data collection;

N_h = total number of eligible arrestees in stratum h.

Expression [5] shows that the number of interviews per facility is either a multiple of 84 or all arrestees available. In reality, ADAM selected all eligible arrestees in small facilities that booked no more than 12 arrestees per day.

The cost model embodied in [4] and [5] is obviously discontinuous. Theoretically, there is no single optimal design when the cost function is not differentiable. On the other hand, as long as an interviewer was assigned to a facility, it was cost effective to complete as many interviews as possible. The resulting sample allocation is typically disproportional in the number of interviews. In particular, the sample allocated to small facilities was greater than their share under proportional allocation, even though the unit cost per interview was higher in these facilities. However, since this allocation minimizes the variance per stratum with a given site budget, it is the most efficient allocation under the cost structure.

III. Errors and Costs Under Two-Stage Designs

ADAM used a two-stage design when there were more than four booking facilities in a site. In this case, it was generally impractical to conduct data collection in all facilities. The two-stage design involved sampling facilities in the first stage and sampling arrestees within facilities in the second stage. Under the two-stage design, each booking facility was viewed as a cluster or a primary sampling unit (PSU) rather than a stratum. The first stage sample was typically stratified and selected with probability proportional to the size of each facility, where size was measured by the total number of arrestees booked in a given period of time. The second stage sample of arrestees was then selected systematically from the sample facilities.

The error and cost models are more complicated under the two-stage design. Ignoring the finite population correction factor in both stages, we can express the general error model as

$$V(\hat{\bar{Y}}) = \sum_{l=1}^{L}\left(\frac{S_{al}^2}{a_l} + \frac{S_{bl}^2}{a_l \bar{b}_l}\right) \quad [6]$$

where

l = first stage sampling strata;
S_{al}^2 = between cluster population variance of stratum l ;
a_l = first stage sample size of stratum l ;
S_{bl}^2 = within cluster population variance of stratum l ;
$\bar{b}_l$ = average second stage sample size of stratum l .

We further assume a cost model of the form

$$C_{cl} = C_0 + \sum_{l=1}^{a_l}\left(C_{1l} a_l\right) + \sum_{l=1}^{a_l}\sum_{k=1}^{b_{lk}}\left(C_{2lk} b_{lk}\right) \quad [7]$$

where

C_{1l} = total fixed costs per facility within stratum l ;
b_{lk} = number of interviews in facility k within stratum l ;
C_{2lk} = cost of sampling and interviewing a single arrestee from facility k within stratum l .

Assuming that the facilities are of equal sizes and that both the facilities and the arrestees are selected by simple random sampling, the optimal cluster size for stratum l is given by

$$b_l = \sqrt{\frac{C_{1l}}{C_{2lk}} \frac{(1-\rho_l)}{\rho_l}} \quad [8]$$

where ρ_l is the intra-cluster correlation coefficient of stratum l (Levy and Lemeshow, 1999) .

Expression [8] provides guidance on the optimal sample size per facility under simplified assumptions. The true cost model under the two-stage design, however, is again a direct function of the number of interviewers rather than the number of interviews. The cost model may be expressed as

$$C_{cl}^{'} = C_0 + \sum_{l=1}^{a_l}\left(C_{1l} a_l\right) + \sum_{l=1}^{a_l}\sum_{k=1}^{b_{lk}^{'}}\left(C_{2lk}^{'} b_{lk}^{'}\right) \quad [9]$$

where

$b_{lk}^{'}$ = number of interviewers in facility k within stratum l ;
$C_{2lk}^{'}$ = cost per interviewer in facility k within stratum l .

Again, the total cost is determined by the number of interviewers rather than the number of interviews completed. The most efficient sample allocation among facilities is to complete as many interviews as possible per facility.

IV. Conclusion

Practical constraints often make it difficult to design cost-effective probability samples, and

compromises have to be made between scientific rigor and practicality. Given the ADAM cost structure, the most efficient sample allocation is to maximize the number of interviews from all facilities in the sample.

ADAM was not designed for national estimates, so sample allocation among sites was not a statistical optimization problem. NIJ decided to make the final allocation roughly proportional to the number of arrestees per site, while maintaining a minimum sample for small sites and a maximum sample for large sites. There was also a sample allocation issue within facilities. This allocation was approximately proportional although this could vary across facilities.

Reference

Cochran, W.G. (1977). *Sampling Techniques,* Third Edition, John Wiley & Sons Inc.

Levy, Paul S. and Stanley Lemeshow (1999). *Sampling of Populations,* Third Edition, John Wiley & Sons Inc.