

## DESIGN AND INFERENCE IN ADAPTIVE SAMPLING

Steven K. Thompson<sup>1</sup>

### ABSTRACT

A basic theorem of sampling says that, under a wide variety of assumed population models, the optimal sampling strategy is an adaptive one. An adaptive design is one in which the selection of units to include in the sample may depend on values of the variable of interest observed during the survey. Examples of adaptive designs include sequential stopping procedures, adaptive stratification and allocation, adaptive cluster sampling, as well as optimal model-based designs. In adaptive cluster sampling, whenever the variable of interest for a unit in the sample satisfies a specified condition, neighboring units are added to the sample. For example, in a survey to estimate the abundance of a rare, clustered animal species, whenever high abundance is encountered additional observations are made at spatially neighboring sites. Conventional estimators would be biased with such a design but simple design unbiased estimators are available. Observational studies---such as the use of catch per unit effort in commercial fisheries management---can also contain adaptive features that may bias the conventional estimates.

Adaptive procedures are also useful in surveys of hidden or hard-to-access human populations. For example, in a study of the behavior of injection drug users, whenever self-reported users are encountered in the sample, social links are followed to add others to the sample. With social network relationships replacing spatial relationships the problem is viewed as sampling in a graph. Link-tracing procedures appear to be essential in such studies in order to obtain samples large enough to study. Inference from the sample to the larger population of interest must take the design or graph structure into account, however. When the statistical inference problem is ignored---as has traditionally been the case---ordinary summaries of sample quantities can be misleading. This result is illustrated with an example.

### RÉSUMÉ

Un théorème de base en échantillonnage dit que, sous une grande variété de modèles présumés pour la population, la stratégie optimale d'échantillonnage est une stratégie adaptative. Un plan adaptatif est un plan tel que la sélection des unités à inclure dans le sondage peut dépendre des valeurs de la variable d'intérêt observée durant l'échantillonnage. Les exemples de plans adaptatifs incluent des procédures d'arrêts séquentielles, de stratification ainsi que de répartition adaptative, de plans en grappes adaptatifs aussi bien que de plans d'échantillonnage optimaux sous le modèle. En échantillonnage de grappes adaptatif, si la variable d'intérêt pour une unité dans l'échantillon satisfait une condition spécifiée, les unités dans un certain voisinage sont ajoutées à l'échantillon. Par exemple, pour estimer la quantité d'une espèce rare d'animaux vivant en grappes, chaque fois qu'une grande quantité est observée, une quantité d'unités additionnelles seraient prélevées à des sites environnants. Les estimateurs conventionnels seraient biaisés dans un tel plan, mais des estimateurs sans biais simples basés sur le plan d'échantillonnage sont disponibles. Les unités, par exemple la prise par unité d'effort dans la gestion de la pêche commerciale, peuvent aussi contenir des caractéristiques adaptatives qui peuvent faire en sorte que les estimateurs conventionnels soient biaisés.

Les procédures adaptatives sont aussi utilisées dans les sondages avec les populations humaines cachées ou difficiles à accéder. Par exemple, dans l'étude du comportement des utilisateurs de drogue par injection où les utilisateurs forment l'échantillon sur une base volontaire, les liens sociaux vont permettre d'ajouter d'autres individus à l'échantillon. Le réseau des relations sociales prenant la place des relations spatiales, le problème est considéré comme l'échantillonnage dans un graphe. Le plan est adaptatif lorsque la décision de poursuivre les liens dépend des valeurs de la variable d'intérêt. Dans de telles études, les procédures de poursuite des liens semblent être essentielles afin d'obtenir des échantillons assez grands pour pouvoir effectuer l'étude. Cependant les inférences provenant d'un échantillon à la plus grande population d'intérêt doivent tenir compte du plan ou de la structure du graphe. Quand on ignore le problème d'inférence statistique, comme cela a été le cas traditionnellement, les statistiques habituelles peuvent porter à confusion. Ce résultat est illustré à l'aide d'un exemple.

---

<sup>1</sup>Department of Statistics, Pennsylvania State University, University Park, PA, 16802, U.S.A.

## 1. INTRODUCTION

An adaptive design is one in which the procedure for selecting units to include in the sample may depend on values of the variable of interest observed during the survey. For example, in a survey of a rare animal species, whenever sufficient abundance of the animals are encountered, additional observations may be made at neighboring sites. Similarly, in an epidemiological study of a rare, contagious disease, whenever a person in the sample is found with the disease, socially linked people may be added to the sample. Examples of adaptive designs include sequential stopping procedures, adaptive stratification and allocation, adaptive cluster sampling, as well as optimal model-based designs (cf., Thompson and Seber 1997).

Motivation for adaptive sampling is provided both by the realities of survey situations in which the population characteristics of interest are rare, clustered, or otherwise difficult to sample by conventional means, and by theoretical results showing that under certain conditions gains in efficiency can be expected through adaptive procedures. A basic theorem of sampling says that, under a wide variety of assumed population models, the optimal sampling strategy is an adaptive one (Zacks 1969, Thompson 1988, Thompson and Seber 1996). The basic idea is that, part way through a survey, if one can observe the variables thus far measured, then the conditional distribution of the unobserved part of the population given the data observed so far provides a guide for selecting the rest of the sample to best effect.

## 2. ON OPTIMAL AND PRACTICAL STRATEGIES

The theoretically optimal sampling plan may not be the most practical to implement, since it is heavily model-based and may require an unrealistic amount of prior knowledge about the population as well as quick access to measurements during the survey and complex computations during or after the survey. Design-based adaptive strategies such as adaptive cluster sampling (Thompson 1990) and certain types of adaptive allocation plans (Thompson, Ramsey, and Seber 1992) offer one approach to achieving a degree of simplicity and robustness. With these strategies, design-unbiased estimators are available that are extremely simple to compute. Further, the unbiasedness of the estimators is based solely on the design and does not rely on any assumptions about the population itself. From a purely design-based approach no optimal strategy for every

possible population configuration exists (Godambe 1955). However, for some types of populations the design-based adaptive strategies tend to have lower mean square error than comparable conventional strategies.

In adaptive cluster sampling, whenever the variable of interest for a unit in the sample satisfies a specified condition, neighboring units are added to the sample. For example, in a survey to estimate the abundance of a rare, clustered animal species, whenever high abundance is encountered additional observations are made at spatially neighboring sites. Conventional estimators would be biased with such a design but simple design unbiased estimators are available. For estimating the abundance of rare, clustered populations, adaptive cluster sampling can produce substantial gains in efficiency relative to conventional designs of equivalent sample size.

## 3. SAMPLING IN A GRAPH

Adaptive procedures are also useful in surveys of hidden or hard-to-access human populations (Thompson 1997). For example, in a study of the behavior of injection drug users, whenever self-reported users are encountered in the sample, social links are followed to add others to the sample. With social network relationships replacing spatial relationships the problem is viewed as sampling in a graph (cf., Frank 1977a,b, 1978a,b, 1979, Frank and Snijders 1994). Since the procedure for selecting the sample depends on node and link values that are only observed as the sample is selected, these link-tracing designs are adaptive. Link-tracing procedures used in studies of hidden and hard-to-reach human populations include various forms of snowball sampling (Goodman 1961), network sampling (Birnbaum and Sirken 1965), chain referral and random walk designs (Klov Dahl 1989). Link-tracing procedures appear to be essential in such studies in order to obtain samples large enough to study. Inference from the sample to the larger population of interest must take the design or graph structure into account, however. When the statistical inference problem is ignored---as has traditionally been the case---ordinary summaries of sample quantities can be misleading.

The following example illustrates a link-tracing strategy in which the design-unbiased estimators of adaptive cluster sampling can be used. The unbiased estimates are contrasted to the conventional sample mean or expansion estimators, which are biased with the

link-tracing selection procedure.

### 3.1 Example

Consider a survey of drug use in a population of 1000 people. The variable of interest is amount spent in the last week on the drug and the object of the survey is to estimate the total amount spent during that period by the population or, equivalently, the mean amount spent per person. An initial sample of 100 people is selected using random sampling without replacement. Drug use is relatively rare in the population, and of the 100 people, only 6 people report any drug use at all. The values reported (in dollars) are 5, 15, 7, 30, 3, and 2, with the other 94 initial respondents reporting zero.

Now to obtain a larger sample of users the investigators will follow social links whenever 10 dollars or more is reported spent. So whenever a respondent reports \$10 or more he or she is asked to name close social contacts (not necessarily drug use contacts), and those linked people are added to the sample. The person who reported spending 15 is asked and names one contact who, when interviewed, reports spending 25. This added person in turn reports two additional people.

But each of those two people reports spending zero, so they are not questioned on their contacts. The person in the initial sample who spent 30 reports two new people, one who spent 100 and one who spent 0; he also reports the person already in the initial sample who spent 7. The added person who spent 100 reports two new people, reporting 20 and 9. The added person who spent 20 is questioned but reports no contacts other than the person already in the sample who had reported him.

Thus, starting with an initial sample of 100 people, the link-tracing design leads to a final sample of 107 people. The naive sample mean of amount spent per person is

$$\bar{y} = (5+15+7+30+3+2+25+100+20+9)/107 = 216/107 = 2.019$$

or just over 2 dollars per person. The conventional expansion estimator of the population total is

$$N\bar{y} = 1000(2.019) = 2019$$

so that the conventional estimate of the size of this underground economy that week is over 2019 dollars.

The final sample contains 10 people who reported any use at all, so the ratio of dollars spent to users in the sample is

$$216/10 = 21.60$$

giving almost 22 dollars per user.

However, these conventional data summary statistics are not unbiased estimates of the corresponding quantities for the population, because of the way the sample was selected. Unbiased estimates for this situation are provided by the design-unbiased estimators of adaptive cluster sampling.

Estimation in adaptive cluster sampling takes uses the network structure in the sample. The person who spent 15 and the person who spent 25 together form one network, because with the design if either one is included in the initial sample both end up in the final sample. The three people reporting 30, 100, and 20 together form another network, because inclusion of any one in the initial sample results in inclusion of all three in the final sample. Each of the other people in the sample forms a network of size one.

The simplest of the design-unbiased estimators simply replaces the original value for each unit in the initial sample with the average of the values in its network. For the network of two units, the average is  $(15+25)/2 = 20$ . For the network of three units, the average is  $(30+100+20)/3 = 50$ .

The unbiased estimator of the mean amount spent per person on drugs in the population is

$$\hat{\mu}_1 = (5 + 20 + 7 + 50 + 3 + 2)/100 = .87/100 = .87$$

so that the unbiased estimate is 87 cents spent per person in contrast to the naive estimate of over two dollars.

An unbiased estimate of the total amount spent in the population is given by the expansion

$$\hat{\tau}_1 = 1000(.87) = 870$$

in contrast to the naive estimate of over 2000 dollars.

There were 6 users in the initial sample, so an unbiased estimate of the number of users in the population is  $100(6)/100 = 60$ . The ratio of unbiased estimates gives

$$870/60 = 14.50$$

or \$14.50 spent on average by each user in the population, in contrast to the naive estimate of almost \$22.

Another type of design-unbiased estimator from adaptive cluster sampling is only slightly less simple to compute and in empirical studies tends to be more efficient than the first. The second estimator divides the total value of a network by the probability that network was intersected by the initial sample, for each network

intersected. In this example, for a network of one person, the intersection probability is simple the probability the person is included in the initial sample, or  $n/N=1$ . For a larger network, the probability of intersection is the probability that one or more of the units in the network are included in the initial sample. This is readily computed as one minus the probability that the initial sample completely misses the network. The computation is straightforward and involves calculating the number of ways to choose the initial sample from the units not in the network. For the network of two people the intersection probability is .19 and for the network of three people it is .27. The second unbiased estimate of the total amount spent is

$$\hat{t}_2 = (5/.1) + (40/.19) + (7/.1) \\ + (150/.27) + (3/.1) + (2/.1) = 936$$

The estimate of total of \$936 in the hidden economic activity is similar to the other unbiased estimate, but again is in contrast to the naive estimate.

The second unbiased estimate of the population mean is

$$\hat{\mu} = 936/1000 = .934$$

or about 94 cents per person.

An unbiased estimate of the number of users in the population is obtained from this method by using as the variable of interest for each person the indicator variable which is one when reported amount spent is greater than zero. The unbiased estimate is

$$(1/.1) + (2/.19) + (1/.1) + (3/.27) + (1/.1) + (1/.1) = 62$$

users in the population. The ratio of unbiased estimates is

$$936/62 = 15.10$$

or about \$15 per user, again in contrast to the naive figure of about \$22.

#### 4. THE EFFECT OF THE DESIGN ON INFERENCE

Design-based sampling strategies have the advantage that properties such as design unbiasedness do not depend on model-based assumptions about the population. However, these properties are entirely dependent on the sampling design being carried out as

specified. In some situations, for example in studies of HIV transmission in relation to drug use, it is very difficult to select the sample by any well-specified means.

In that case the more useful approach may be to consider a model for the population, such as a stochastic graph model, and base inference on the model.

Model-based estimation or other inference is simplified if the design can be ignored when making the inference. The fact that the sampling design does not affect the value of a Bayes estimator in survey sampling when the design does not depend on values of the variable of interest outside the sample was noted by Basu (1969). Rubin (1976) gave exact conditions for a missing data mechanism---of which a sampling design is one example---to be relevant in frequentist and likelihood-based inference. For likelihood-based methods such as maximum likelihood and Bayes methods, the design is ignorable if the design or mechanism does not depend on values of the variable of interest outside the sample or on a parameter in the distribution of those values. For frequency-based inference such as unbiased estimation, however, the design is relevant if it depends on values of the variable of interest even in the sample. Scott (1977) showed that the design is relevant to Bayes estimation if auxiliary information used in the design is not available at the inference stage. Sugden and Smith (1984) extended results on when the design is relevant to predictive inferences.

With adaptive sampling designs the selection procedure deliberately seeks to take advantage of observed values of the variable of interest. The well-established adaptive designs are ignorable, since the procedure depends only on values in the sample. However, with some sampling procedures surprising dependencies on unobserved values can occur, particularly with link-tracing designs in the graph setting, and ignorability of the design can be sensitive to exactly what data are collected (Thompson and Frank 1977).

#### ACKNOWLEDGEMENTS

Support for this research was provided by the National Science Foundation, grant DMS-9626102, and the National Institutes of Health, National Institute on Drug Abuse, grant RO1 DA09872-01A2

#### REFERENCES

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhya A*, 31, 441--454.

- Birnbaum, Z.W., and Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics, Ser. 2, No.11*. Washington:Government Printing Office.
- Frank, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference, 1*, 235-264.
- Frank, O. (1977b). Estimation of graph totals. *Scandinavian Journal of Statistics, 4*, 81-89.
- Frank, O. (1978a). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics, 5*, 177-188.
- Frank, O. (1978b). Sampling and estimation in large social networks. *Social Networks, 1*, 91-101.
- Frank, O. (1979). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research*, edited by P.W. Holland and S. Leinhardt. New York: Academic Press, 319-347.
- Frank, O., and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics, 10*, 53-67.
- Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society B, 17*, 269-278.
- Goodman, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics, 32* 148-170.
- Klov Dahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In M. Kochen, ed. *The Small World*, Norwood, NJ: Ablex Publishing, 176-210.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.
- Scott, A.J. (1977). On the problem of randomization in survey sampling. *Sankhya C, 39*, 1--9.
- Sugden, R.A., and Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika, 71*, 495-506.
- Thompson, S.K. (1988). Adaptive sampling. *Proceedings of the Section on Survey Research Methods of the American Statistical Association, 784-786*.
- Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association, 85*, 1050-1059.
- Thompson, S.K. (1997). Adaptive sampling in behavioral surveys. In Harrison, L., and Hughes, A. eds., *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*. NIDA Research Monograph 167. Rockville, MD: National Institute of Drug Abuse, 296-319.
- Thompson, S.K., Ramsey, F.L., and Seber, G.A.F. (1992). An adaptive procedure for sampling animal populations. *Biometrics, 48*, 1195--1199.
- Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: Wiley.
- Zacks, S. (1969). Bayes sequential designs of fixed size samples from finite populations. *Journal of the American Statistical Association, 64*, 1342--1349.