DISCUSSION

Charles H. Alexander, U. S. Bureau of the Census
Demographical Statistical Methods Division, Washington, D.C. 20233

KEY WORDS: Sample surveys, conditional inference, superpopulation model

INTRODUCTION

My father used to tell me stories about Morris Hansen when I was a kid. His main point was that the reason Hansen was able to accomplish so much was that he started out his career with some important cost-saving innovations, such as using sampling in the census, so it was easy for his Budget and Finance Division Chief (my father) to get him money for other things he wanted to do. Somehow the moral was that I should save my allowance and not spend it on candy, so I started out with mixed feelings about Morris Hansen. But in later years, my father often spoke about his admiration for Hansen and the sense of excitement and accomplishment at Census during the Hansen years, which Professor Smith has mentioned. So I'm personally very grateful to be able to participate in this program in honor of Morris Hansen.

I also started out with mixed feelings about T.M.F. Smith, when people told me that he had proved that sampling and weighting are a total waste of time. However, when I first heard Professor Smith speak, some years ago, I realized that the reports I had heard about his views on the foundations of inference from sample surveys had been badly oversimplified. Although he did not come down on the side of us randomizers, his comments showed a remarkable understanding of what real-life survey work is all about, and got right at the crucial issues.

I'm glad the paper did not dwell on the history of the conflict, so I don't feel obliged to do a complete enumeration of the randomizer's arguments. Over the years, Hansen and his colleagues, as well as others such as Graham Kalton and Leslie Kish have made a thorough presentation of the issues from the survey sampling point of view. Prof. Smith's paper (like his earlier work) takes these arguments into account. He has been a major contributor to what he calls "the coming together of minds" on the usefulness of both models and randomization with known selection probabilities for design and estimation.

I'm specifically relieved that I don't have to try to speak for Morris Hansen and restate his principles to make sure they're presented fairly. The paper gives a balanced presentation, mostly letting Hansen and colleagues speak for themselves with an admirably representative selection of their writings. (Of course, from my point of view, a random sample of paragraphs would also be called unbiased.)

Prof. Smith's paper gives a good review of recent research and open issues in survey work. I thought this was right on the mark, except that I would have nominated either response error or undercoverage as the Achilles heel, rather than the somewhat better-understood problem of missing observations.

To me, though, the most exciting part of the paper was the sketch of an alternative framework for randomization inference, inspired by Fisher consistency and the two references by Robinson. I'll have a few more comments about that. But mostly I want to talk about Prof. Smith's idea of inference as a "social phenomenon", and his dramatic conclusion.

It has always struck me that peoples' arguments about the merits of randomization inference for surveys are influenced by which survey they are thinking about. Prof. Smith's discussion points out that an equally important consideration is who will be making the inference. Is the statistician going to be making a personal inference, which non-statisticians are to accept as expert opinion, or is the statistician's role to provide information for non- statisticians to make their own inferences? I think Hansen always assumed the latter.

Most of what I know about Morris Hansen's philosophy comes from having labored in the garden which he planted at the Census Bureau. The Demographic Statistical Methods Division, where I work, has a distinctive approach to the practice of household surveys, which we attribute to Morris Hansen, and others whom he strongly influenced. I want to describe some practical aspects of our "social phenomenon" of randomization inference. Keep in mind that this is my interpretation and for the most part does not come directly from Hansen or his writings.

INFLUENCE OF THE CPS PARADIGM ON THE DEVELOPMENT OF SURVEY INFERENCE

First, the HMT article, along with much of the debate about randomization inference over the last two decades, focusses on a model which describes a typical establishment or business survey. However, I conjecture that Hansen's paradigm throughout has been household surveys, particularly the U.S. labor force survey, now known as the Current Population Survey (CPS). Some relevant features of CPS are:

a. There is no useful design information available for all individual households except at census time. Information is available from the previous census for PSUs or city blocks, but this information is out-of-date, not highly correlated with labor force characteristics, and much of it is already used in stratification. For those of us who started out on household surveys, it is hard to accept that the essence of survey inference is how to predict Y based on a strong relationship with a well-known X.

b. Things are further complicated by the fact that you don't have a complete list of households, nor even know the population size. In fact, some of the cleverest inventions of Hansen and his colleagues are operational methods for selecting a sample indirectly without a complete list of the target population, and still being sure that in principle every household has a positive probability of selection and that you know the probability for any household which you actually select. The most familiar examples are the methods used for selecting area samples, but similar techniques are used for sampling from census lists and building permit offices.

I think it is natural for someone faced with these two problems to lean towards a randomization approach.

My next two points may help explain why the sample survey tradition is relatively lacking in attention to conditions for the normal approximation to apply:

c. CPS's primary variables of interest, unemployment and labor force participation, are binary.

d. National CPS estimates have very small between-PSU variance for unemployment and labor force participation. This is true for national CPS estimates with the current design, and I would guess this was true in the past, even when the survey had fewer PSUs.

For the CPS, the population values of greatest interest are very well-behaved. Therefore the adequacy of the normal approximation is probably not of practical importance. For other surveys, or even CPS estimates for individual states, this may not be the case.

While I'm on this topic, let me say that I found Prof. Smith's proposed use of finite-population consistency to be very appealing. I think it does correspond more closely than asymptotic consistency to the intuitions we samplers use in coming up with estimators. I've been uncomfortable with the HMT asymptotic approach, mainly because the Brewer formulation which is used in HMT, starts the sequence with your sample (and population) and goes on from there, so there's no reason to hope the limit of the infinite sequence says much about your sample. So I'm looking forward to the forthcoming paper by Smith and Scott. My one question for them is whether Fisher consistency implies anything about how close your estimator is likely to be to the population value of interest, or has other obviously desirable properties.

I add two more features of CPS, whose implications I won't address:

e. For CPS, systematic sampling within PSUs is reasonably modelled by simple random sampling of clusters, within large strata. This is because the sort order of households is based on variables thought to be noninformative.

f. CPS has many important secondary purposes.

THE "SOCIAL PHENOMENON" OF INFERENCE FROM OFFICIAL SURVEYS

These are the tasks of the survey practitioner relating to inference:

a. Design and select the sample.

b. Develop or choose the estimator.

c. Defend (to "politicians") the choice of the estimator. By "politician," I mean anyone who is hoping to see your survey yield a particular result. If you don't get that result, you'll be encouraged to reconsider your estimator. This is particularly important for a new survey where you can't say "we used the same formula we've always used" or "we couldn't afford to change the computer program."

d. Calculate a confidence interval.

e. Interpret the confidence interval for a general audience. We usually do this in a "Source and Accuracy" statement in the back of our reports.

Let me first focus on tasks c and e, which most directly affect inference. The need to defend your choice of estimator has some subtle consequences on the practice of randomization inference. As Prof. Smith observed, in theory a randomizer can agree a stratifier is relevant, but can ignore it for post-stratification and still have a robust (though inefficient) estimator. Many statisticians have an image of us saying to our survey sponsors, "unfortunately this year we had an unlucky sample and you have an obvious underestimate, but it's not a problem: you equally well might have gotten an obvious overestimate." In practice, we post-stratify (sometimes to administrative data and sometimes to a larger survey) on any stratifier which is obviously relevant.

Ideally, we'd like to choose the estimator before seeing the data, but sometimes we don't realize the need to post-stratify until it's too late. We randomizers undergo a great deal of introspection whenever we have to improve the estimator after seeing the data. Are we sure we're not unconsciously "cooking" the data? ("Cooking the data" refers to trying out lots of estimators on your particular sample in hopes of getting the answer some "politician" wanted.) We ask ourselves "we now see that our sample over-represents central cities, so our national poverty rate is too high, but would we have altered the estimator if we had seen too little sample in central cities?"

This line of thought is not part of the standard theory of inference, but I think it's not a trivial concern, especially for official statistics. It naturally leads you to consider samples that you might have picked but didn't.

The issue is whether unconscious biases can affect your choice of estimator, just as we all worry that they might affect your choice of sample. If your choice is based on reducing the variance over all possible samples from your design, you can comfort yourself that you should have made the same choice even if you'd observed a different sample, so you can't be cooking the data. (This isn't totally convincing, since it's not clear you would have even asked the question if you'd had a different sample, but at least your freedom to "cook the data" is somewhat limited.)

Our final step in practical inference was explaining the confidence interval to a general audience. This is the sort of thing we say in our Source and Accuracy Statement:

> "Keep in mind that the particular sample we selected was one of many possible samples... Different samples would give different results."

Then we give a statement about how confidence intervals would perform in repeated sampling, something like:

> "If many samples were selected, then for approximately 95% of the samples the confidence interval which would be calculated would contain the result which would be obtained by surveying the entire population."

It sounds like someone's worst nightmare from Freshman Statistics, but it does two things:

a. It gives the readers a concrete image to reinforce the notion that there is uncertainty because of sampling error.

b. It precisely and completely states the fact upon which we expect the readers to base their inferences about sampling error.

As far as nonsampling error, we try to tell the readers as much as we know about our data collection and estimation procedures and how they affect the accuracy of the data. But in the end, we force the readers to make their own "scientific" inferences about nonsampling error. An example of a scientific inference in criminology would be "it seems to me that your survey procedures will tend to miss many drug dealers, and even though you give higher weight to sample persons (mostly non-drug dealers) with matching demographic characteristics, you undoubtedly underestimate gunshot victims." It's clear that in the social sciences, if you have to extrapolate from your survey data, you'd rather do it "statistically" than "scientifically".

This illustrates a difference between how models are used in science and in descriptive surveys. A scientific model's intuitive plausibility is ultimately not crucial, because the model will be used to make predictions which will be tested against observed values. By contrast, the models for nonresponse, etc., in official statistics are typically used without testing to generate "observed" values. These "observations" are often used by social scientists to test their theories. Thus the social scientist is very dependent on the credibility of the survey statistician's models. This highlights the importance of scientific work on survey methods, to test there models, which Prof. Smith reviews.

Returning to sampling error, how well do we communicate uncertainty due to sampling to the people who must make inferences from our survey data? My favorite line from Prof. Smith's paper was "All inferences are the product of human imagination and there can be no absolutely correct method of inductive reasoning." I equally liked this phrase from Smith (1991): "...if variances are used as a broad general guide to accuracy, not as part of a precise inference..." The "imprecision" of the randomization approach to inference presents two main problems for communicating uncertainty to a general audience.

First, randomization theory gives us some impressive facts about 95% of the samples we might have gotten, but there's no compelling response if someone asks "so what?" We can say "either the population value is in this interval or we got an unusual sample," to which the natural reply is "oh, O.K., which is it?" At that point we're stuck, especially if there is evidence that the sample is actually a bit unusual.

Second, there is the technical question of how good the normal approximation is. I agree with Prof. Smith's polite hints that survey practice needs to include more work on verifying this approximation and acknowledging it when we discuss uncertainty. In some cases (e.g., binary variables) I think that with more work we could pin the results down so well that uncertainty about the "95%" value would be no more part of the inference problem than whether the calculator used to add up the figures worked correctly. The HMT article suggests to me that this technical question should be external to the non-statistician's inferential thought process, just like the calculator. I personally think that's sometimes reasonable, but sometimes there's more doubt about the approximation, which we need to communicate better than we do.

But, all in all, I'd say randomization inference can be communicated adequately, if imperfectly, to a variety of readers in a Source and Accuracy Statement. How about model-based inference? Is it possible to communicate, in a brief statement for nonstatisticians, the understanding necessary for them to make a model-based inference about uncertainty due to sampling, including a measure of the impact of uncertainty about the model. I have not seen this done. I'll leave it as a challenge for someone to draft a Source and Accuracy Statement for a general audience from the model-based perspective.

SOME OPEN THEORETICAL QUESTIONS

While I'm on the subject of challenges, I'd like to suggest some open theoretical problems related to our practice. To introduce the first two questions, I need to explain that our estimators are derived as separate solutions to distinct, but sort of nested problems, making sure that the answer to each problem is a weight, and then the weights are multiplied together. These are the problems:

First stage of selection. There is some useful design information (X). This can be viewed as a ratio or regression estimation problem.

Unequal probabilities at second stage. Here there is no useful X for most household surveys. The population size is unknown.

Unplanned subsampling. In several situations, the original design may give a particular interviewer too much work to complete. A subsample of the assignment is retained; the inverse of the subsampling rates is used as a weight.

Adjustment for non-response. This requires models, although response can be modelled as a fixed characteristic if you choose to do so.

Post-stratification for person weights. Since this adjustment partially corrects for undercoverage, some modelling is involved.

Model-based adjustment for household weights. This refers to the principal person weight, which I think is best motivated as maximum likelihood estimation under a reasonable model for undercoverage. (See Alexander (1987, 1989).)

Some of these problems can be addressed with a randomization approach; others require modelling. This leads to my first open question.

a. How are the different stages of survey weighting to be combined conceptually?

Expressing the whole thing as one big likelihood function would be a major challenge. On the other hand, I'm not sure exactly what it means to treat nonresponse and undercoverage as part of the variation over all possible samples. So how to fit all this together? One possible approach is what has been called pseudo-maximum likelihood estimation, where some parameters are replaced by asymptotically consistent estimates and the likelihood is maximized over the other parameters (Gong and Samaniego (1981)). Maybe finite-sample consistency could be used instead.

b. Exactly what meaning do replication (or jackknife) estimates of variance have when applied to our multistage estimators which combine randomization-based and model-based weights?

Prof. Smith concluded that the irreconcilable issue is what to use for the variance. Nowadays we at Census pretty much always calculate variances using a replication (or sometimes jackknife) method. The exact meaning of that variance is somewhat unclear, given the mixture of randomization and models in the estimator, not to mention the omnipresence of systematic sampling. Is our variance estimated conditionally or unconditionally or what?

Recent work by Valliant begins to answer this question. Valliant (1991) shows, for one-stage stratified cluster sampling, that the standard variance methods give asymptotically correct conditional variances for the post-stratified estimators. He contrasts this with the asymptotic unconditional variance of an un-post-stratified estimator, but does not discuss what we randomizers want to look at, which is the asymptotic unconditional variance of the post-stratified estimator.

If you found that hard to follow, the randomizer's view is stated better by Prof. Smith when he says: "I can accept the use of auxiliary information...to reduce the variation in population values, but I now think that the framework for descriptive inference should be the unconditional distribution relating to the original sampling procedure."

c. Exactly where do we draw the line on "all possible samples"?

When I first read Prof. Smith's conclusion, I was troubled by his rejection of conditional randomization inference, in spite of the appeal of conditioning to HMT, Fuller (1981), Rao (1985) and others. My concern was how do we draw the line on "all possible samples"? Isn't there some chance we could have used a different sample size or even a different sample design? Do we need to consider those possible samples?

I now see that I focussed too much on "unconditional" and not enough on "procedural". Hansen's procedural inferences rely on:

1) Knowing based only on the sample design that $E\hat{Y} \doteq Y$, where the expectation of the estimator is taken over some set of possible samples;

2) getting an estimate of Var $(\hat{Y})$ which is approximately independent of $\hat{Y}$;

3) being willing to approximate the distribution of $\hat{Y}$ by a normal distribution.

The crucial condition $E\hat{Y} \doteq Y$ usually requires that the expectation be taken over all the possible samples from some sample design. The one notable exception is post-stratification after SRS. This is why procedural inferences based on the sample design are in general unconditional. That's all there is to it. By insisting on unconditional randomization inference, Prof. Smith has not introduced any "maximum unconditionality principle" which requires us to expand the set of possible samples as far as possible in order to make a proper inference. There is no need to go beyond what we need for these three conditions. So I now don't think that his stance introduces any fatal ambiguities about conditioning. (Sequential sampling procedures, in which the data from earlier sample units influence the selection of subsequent sample units, are another story.)

d. Does the randomization approach apply to superpopulation analysis in some cases?

I'd like to extend Prof. Smith's concluding arguments in favor of procedural descriptive inference for finite populations to cover inferences about "super- populations," in the sense the term is often used in household surveys. Suppose we're interested in whether the relationship between education level and being a victim of crime has changed between last year and this. Many data users insist they are not interested in the actual finite population, but in whether the underlying process of victimization has changed. Here "underlying process" doesn't mean a specific model, but a recognition that being a victim has a certain element of chance--your neighbor's dog doesn't wake up and scare off the burglar--so the analyst can imagine the population's crime rate having been a bit different without any meaningful changes in the underlying state of nature.

Graham Kalton argues in a 1983 paper that you can apply randomization inference to this kind of superpopulation, basically on the grounds that the variation between the large finite population and the superpopulation is ordinarily negligible compared to the variation between the sample and either one. So your procedural descriptive inference may equally well apply to the finite population or a hypothetical larger population from which it was drawn.

I think that's the right way to look at it. We need to distinguish between analysts who really have a superpopulation model and analysts who (to use a term from Kish (1992)), are "population bound," but still want to think of the finite population as a realization of some unspecified "superpopulation phenomenon". I think the concluding section of Prof. Smith's paper supports the argument that the latter analyst should be using randomization inference.

CONCLUSION

In conclusion, I want to emphasize the dramatic conclusions in the paper's last section. Prof. Smith has been one of the leading opponents of Morris Hansen's philosophy of (descriptive) statistical inference, certainly the leading opponent as far as listening with care and understanding to what Hansen was trying to say, but in the end rejecting Hansen's position. Some of us had looked to Prof. Smith for a "reconciliation" of the randomization and model-based approaches, which was to consist of a discovery of some astoundingly robust model, under which the randomizer's procedural inferences were implied by the likelihood principle.

There seemed to be momentum toward such as a reconciliation. Random sampling can now be viewed as a device to give modelers an uninformative design (Sugden and Smith (1984)). Survey weights

turned out to have a model-based interpretation (Smith (1988)). Some radically different prescriptions from the two camps about how to use auxiliary information in specific situations converged substantially. (See Section 4.4 of HMT.) Model-based estimators were developed using "design consistency" to increase robustness (see Kott (1990) and also Little (1983)). Unfortunately, in the end there is no reconciliation between the approaches about what is the variance and what does it mean.

However, there is a reconciliation of sorts in Prof. Smith's conclusion, a reconciliation to Morris Hansen's "real" world in which even the most recognizable reference sets are not homogeneous. In this world, we still don't have a complete explanation of the final leap in the induction from sample to population. But Prof. Smith explains clearly why we should not kid ourselves that it is the same inductive process that lets us make inferences about a precise scientific model.

Sadly, Morris Hansen is not here to welcome Prof. Smith into the fold. Certainly the rest of us in the randomization camp welcome him with open arms. I found his paper to be worth reading and re-reading and I think it will be read and discussed for some time to come.

References

Alexander, C. H. (1987). A Class of Methods for Using Person Controls in Household Weighting. Survey Methodology, 13, pp. 183-198.

Alexander, C. H. (1990). Incorporating Person Estimates into Household Weighting Using Various Models for Coverage. Proceedings of the 1990 Annual Research Conference, Bureau of the Census, pp. 445-462.

Fuller, W. A. (1981). "Comment" on "Ratio Estimator and Estimators of its Variance," Journal of the American Statistical Association, 75, pp. 261-168.

Gong, G. and Samaniego, F. J. (1981). Pseudo Maximum Likelihood Estimation: Theory and Applications. Annals of Statistics, 9, pp. 861-869.

Kalton, G. (1983). Models in the Practice of Survey Sampling. International Statistical Review, 51, pp. 175-188.

Kish, L. (1992). Weighting for Unequal P_i. Journal of Official Statistics, [to appear]

Kott, P.S. (1990). The Design Consistent Regression Estimator and its Conditional Variance. Journal of Statistical Planning and Inference. 24, pp. 287-296.

Valliant, R. (1991). Post-stratification and Conditional Variance Estimation. Presented at the 1991 Convention of the American Statistical Association.