DISCUSSION

Carl Erik Särndal, Université de Montréal

I wish to add my congratulations to Hansen, Madow and Tepping (HMT) on what I think is an excellent statement of the position taken by the classical school in relation to some of the new developments in inference in survey sampling. During the last 15 years, when the foundations of survey sampling have been debated, it is rarely that we have seen a statement from statisticians who pioneered in survey sampling that is as extensive as the one we are discussing today.

The title of this session is "Foundations of Survey Sampling". HMT have chosen to dwell extensively on one aspect of this topic namely the differences between the Probability-sampling theory approach and the Model-dependent approach. I too have experienced this as a central issue, although, when I think of "foundations of survey sampling", a range of other topics come to mind. HMT do a good job of further clarifying the essentials of the two approaches, which have also been called "design-based" and "model-based", respectively.

To me, they represent two different attitudes or frameworks for thought. The basic propositions are different, the conclusions are often (in spite of what we read in this paper) surprisingly similar. The probability-sampling thinking is centered around the randomization probabilities of the sampling design and the inferences that flow from them, while the model-based thinking is an adaptation to survey sampling of things statisticians do when faced with estimation of parameters in experimental models or cause systems. (Incidentally, this means that even the simplest of concepts, such as unbiasedness, has two different meanings which is bound to lead to some unfortunate confusion to students of the area.)

My first comment relates to the question: Why should survey samplers repress their use of statistical models when statisticians in all other branches of statistics feel free to benefit from models; in short: Why should survey samplers be different from everybody else. In recent discussions of survey sampling this question has often been heard; clearly the HMT paper is relevant in trying to formulate an answer. I remember how about 10 years ago the late Allan Birnbaum, knowing of my interest in survey sampling, asked me roughly this question. I think he was surprised at what to him seemed an inefficient approach, namely, in essence to make inference without systematic use of models. This was the reaction of a man who had thought deeply about the foundations of general statistical inference, as opposed to the foundations of inference in survey sampling. This latter topic was at the time diffusely defined or almost nonexistent since only a handful of people had given serious thought thereto.

Several times recently, I have had occasion, at seminar presentations and in the classroom, to discuss in what I consider a fairly balanced way the main features of design-based (= probability-sampling) inference on the one hand and model-based inference on the other. When you present the two arguments side by side, giving people the possibility of direct comparison, you often find that the model-based thinking has the more direct appeal, that it is preferred on intuitive grounds. One reason for this is I think that in any randomly assembled group of statisticians, the people who have genuine appreciation of the exigencies of survey sampling practice, as we know it from classical texts, are simply outnumbered by those who feel more at home with direct model reasoning. Survey sampling, to the average statistician, is not a wellknown field. And with model-based thinking, it is easy to see why a certain estimator is preferred in a given situation. As a beginning student, I can remember how hard it was in reading pioneering sampling texts, such as those of Cochran and of Hansen, Hurwitz, Madow, to understand the reasons behind the more complex procedures.

Today, I find that many of these difficulties of comprehension are resolved by a model oriented frame of reference, and this does not prevent me from appreciating the importance of the arguments advanced, as in this paper, by proponents of the probability-sampling approach.

Every survey sampling statistician would probably in the long run arrive at a philosophy where probability sampling elements and model-based elements are mixed, but where emphasis varies; HMT are no exception to this.

Now in today's paper, HMT make it explicit that models are present in many facets of their approach, somewhat in the background perhaps, but models are very definitely used. At the same time, HMT tell us why a more courageous, more direct appeal to models is not possible, in their opinion, to meet the goals of large scale surveys. Abstracting from this rich paper, we conclude that key elements in their "conservative" attitude are that large sample inferences must be independent of an assumed model, that consistency (in their sense of the term) is an absolute requirement, that "best" estimators play a subordinate role, that more important than maximum efficiency are issues relating to time and cost.

There are some fine distinctions here of which we must be aware. HMT make it very clear that their reasoning applies when we are estimating certain elementary characteristics of the population being surveyed, principally means and totals. In fact, 95% of the literature on survey sampling seems to be concerned with this target of estimation, the sum or the mean of a finite set of numbers, which in a wider perspective seems like a very limited problem. However, it happens to be one that is extremely important to statistical agencies producing large sample estimates at national or regional levels. Here, errors in estimation can be claimed to be so costly that it becomes difficult to argue with the proposition

that an extremely prudent approach, independent of population form, must be maintained at all times, and this in what HMT tell us we have to do.

In many other practical survey sampling problems, it is fruitful or even necessary to bring in a model, for example, in connection with non-sampling errors, or when we consider a cause system, perhaps involving regression equations. The objective may be to estimate parameters of the cause system itself, or to predict the mean of a finite universe which we assume to be the realization at some future moment in time of the cause system. HMT concede that in such a situation we have no choice but to use a model-dependent approach. Concerning estimation of regression parameters, there is disagreement in the literature regarding the appropriate role of randomization probabilities in the estimation procedure. HMT refer briefly to the work of Kish in this area, in which there are also other points of view.

"Methods of data analysis for sample surveys" is a topic which has been discussed elsewhere at these meetings. This is something many statisticians and data analysts would like to know more about, and survey sampling specialists ought to have many imputs to make in such a discussion. But as long as "survey sampling" is taken to mean "estimation of the finite population total (or mean)", survey samplers are not going to be too helpful in a broader perspective, that is, to data analysts who want to know how to account in their analysis for the complex sampling designs often used in gathering the data. HMT's paper is not of much help in this area. It seems clear to me that the model-based framework, being of wider scope, will prove superior in the development of this area; perhaps this is where model-based samplers ought to concentrate their efforts instead of attempting to reassess methods for estimating $\Sigma_1^N y_k$; after all, the "old" methods for doing this work well.

My next point concerns the role of empirical data examples in this type of discussion. HMT start their paper by a long numerical example showing that a model-dependent estimator goes wrong (becomes biased) when the assumed model is false. (Incidentally, requiring a large population and a large sample situation to prove the point, HMT's example will hopefully inspire additional realistic empirical research in survey sampling. By unrealistic I mean those all too often seen empirical comparisons executed for populations of size $N = 10$ and the like.) It is an excellent example, but it raises same questions. Personally, I find that when the complexities are so many, it is difficult to construct an example that is perfectly equitable to both sides. HMT take three probability-sampling estimates and compare them with two model-dependent estimates. Where do these come from? Each of the two model-dependent estimators is derived from simple model, which happens to be false, hence the bias. (Other model-dependent alternatives should have been considered!) The three probability-sampling estimators are suggested; here some experience in sampling is needed to know a priori that they will work out about equally well in the given situation.

The probability sampling estimators (but not the model dependent ones) benefit from knowing something about the population, namely, its shape. The weighted-mean estimator becomes more efficient by sampling more heavily the strata containing large units. Here, with each stratum contributing 10% of the total sample, approximately optimum allocation requires about ten times as large a sampling fraction in the stratum of the largest units as in the stratum of the smallest units. The example is constructed so that the conditions are ideal for the weighted-mean estimator. In spite of this, and even though the assumed models are wrong, the model-dependent estimators perform well for small samples.

In a sense it is necessary to pursue the example further. After all, if we really believe in (approximately) a regression through the origin, the stratum of the largest units should be sampled much more heavily, without necessarily reaching the extreme where all of the sample comes from the stratum of the largest units. What happens in a comparison between the five estimators when the stratum of the largest units contributes say 90% of the total sample with the remaining 10% distributed among the other 9 strata?

The negative bias of, say, the Model I estimator would persist, probably become larger, but its variance should be very small. But the weighted-mean estimator, for example, would have a large variance. Would there again be a sample size n at which the weighted-mean estimator "catches up" and becomes better in an MSE sense? Would this n be so large that for all practical sample sizes the model estimator is preferred on the basis of smaller MSE? For a full understanding of the situation one should show what happens under conditions of sampling other than proportional and approximately optimal allocation.

The HMT paper deserves careful study by anyone interested in the Foundations of Survey Sampling.