Journal of Official Statistics
Vol. 1, No. 1, 1985, pp. 49–63
© Statistics Sweden

How Survey Methodologists Communicate

Carl-Erik Särndal[1]

Abstract: A number of increasingly popular survey sampling terms are built on the distinction between model-based and design-based, for example, is "design-based inference" as opposed to "model-based inference". These two concepts appear to have clearly defined meanings in the literature. The distinction is also often applied, but not without ambiguity, to nouns such as approach, theory, estimator and others. We examine some of the resulting terms and suggest interpretations in cases where the meaning is not automatically clear.

Key words: Survey sampling, design-based, model-based, inference, confidence intervals, (generalized) regression estimator.

1. Introduction

This paper deals with some basic conceptual issues, but does not so much concern the foundations of survey sampling as the language currently in vogue among survey methodologists.

In the summer of 1977, at the 7th Nordic Conference on Mathematical Statistics, I lectured on "Design-based and model-based inference in survey sampling". (The notes were subsequently published; Särndal (1978).) Given the time elapsed, the title now strikes me as interesting. The distinction between "design-based" versus "model-based" was, I believe, a natural linguistic byproduct of the sometimes intense discussions that specialists in the foundations of survey sampling were involved in during the 1970's. A short and descriptive vocabulary was needed to convey a basic distinction, and so the terms emerged.

The distinction caught on – apparently it filled a need – and today "design-based" and "model-based" are popular modifiers of a number of nouns; they are increasingly seen in important places such as titles of papers, lists of key words, and titles of sessions at professional meetings.

"Inference" is far from the only noun that is qualified, in the current sampling literature, as being either "model-based" or "design-based". The distinction is often applied to other words such as approach, theory, viewpoint, standpoint, framework and analysis. The exact meaning of the resulting terms is not always clear, in my opinion. The most common of all applications of the distinction is perhaps in qualifying the word "estimator". As I argue in Section 8, "design-based estimator" and "model-based estimator" are particularly unfortunate and ambiguous terms. In addition, writers use related terms such as "model-dependent" (approach, estimator, etc.) and "model-free" (approach, estimator, etc.) and even "design-free".

Those wishing to see a number of these terms in actual use are referred to the recent important paper by Hansen, Madow and

[1] Professor, Université de Montréal, Canada. This research was supported by the Natural Sciences and Engineering Research Council of Canada. The author is also grateful to colleagues at Statistics Canada and at Statistics Sweden for discussions that in part inspired him to these thoughts. The perceptive comments of two referees helped to improve the original manuscript.

Tepping (1983), and in particular to the long discussion that follows that article. We should keep in mind that there do not exist generally agreed upon definitions for some of the terms. Rather, they are used informally, for rapid (but often imprecise) communication. When terminology is not firmly rooted, hesitation and risks of misunderstanding arise. For example, in commenting on the terms "model-based inference" and "model-dependent inference" used by Hansen, Madow and Tepping (1983), Little (1983) states "I dislike this terminology, and shall use the terms model-based design, model-based estimation, model-based inference to describe the use of models ...". If no eyebrows are raised when someone dislikes certain scientific terms and promptly proposes his own "improved" terms, it is, of course, because no generally accepted terminology exists, and because the scientific field is in a state of development. I hope that the survey sampling literature is at the point of converging on a terminology that we can all accept. If not, discussions of basic issues in survey sampling will continue to be difficult to follow both by theoreticians and practitioners.

The comprehensive paper by Kalton (1983) on utilization of models in surveys is another example of an article with frequent use of terms that involve the distinction between design-based and model-based. Other references of interest in this connection are Rennermalm (1980), Smith (1981), Holt, Smith and Winter (1980), Little (1982), Sundberg (1983), Wright (1983), Särndal and Wright (1984).

I criticize no one in this connection, except myself, for a certain inconsistency. Firmly rooted definitions have largely been lacking, and I notice that my own usage of these terms has not been entirely consistent, as seen by comparing Särndal (1978), (1980), (1981), (1982), (1984). This consideration underlies the remarks in this paper.

For sake of illustration, I have chosen to discuss the design-based versus model-based distinction as applied to three concepts, namely, "inference" (Sections 2–5), "approach" (Section 6) and "estimator" (Sections 7–8). Except in the case of "inference", the resulting terms are not without ambiguity and the risks for misunderstanding are considerable.

2. Design-based confidence intervals

As is customary, we mean by "inference" a statement made about an unknown population quantity, in terms not of full certainty but of probability. In inference by means of a confidence interval, the uncertainty is expressed by a "confidence level" of, say, 95 %; in the case of inference through a hypothesis test, by a "significance level" of, say, 1 %. In the discussion below, we concentrate on confidence intervals for the finite population mean $\bar{y} = \Sigma_1^N y_k/N$. (The population, denoted U, consists of N units labelled $k = 1,\ldots,N$.) If $\hat{\bar{y}}$ is an estimator of $\bar{y}$, then most sampling statisticians will find it natural to calculate a confidence interval as

$$\hat{\bar{y}} \pm 1.96\sqrt{\hat{V}_p(\hat{\bar{y}})} \quad . \tag{2.1}$$

The constant 1.96 guarantees an approximately 95 % level of confidence, in sufficiently large samples, and $\hat{V}_p(\hat{\bar{y}})$ is an estimator of $V_p(\hat{\bar{y}})$, the variance of $\hat{\bar{y}}$ "over repeated samples" drawn with the sampling design denoted p. It is also popular among survey statisticians to express the precision of the estimate $\hat{\bar{y}}$ in terms of the coefficient of variation,

$$\mathrm{cv}(\hat{\bar{y}}) = \sqrt{\hat{V}_p(\hat{\bar{y}})} / \hat{\bar{y}} \quad .$$

For example, one may be satisfied that the estimate is sufficiently precise if $\mathrm{cv}(\hat{\bar{y}}) \leqslant .10$.

We need only think for a moment to realize that the use of the interval (2.1) rests on at least two additional assumptions. One is that $\hat{\bar{y}}$ is an approximately unbiased estimator of $\bar{y}$. Otherwise, if in large samples the bias of $\hat{\bar{y}}$ does not approach zero while the variance does, then the interval (2.1) will have poor coverage properties: The probability of covering $\bar{y}$ will be essentially zero in large samples. We therefore assume that $\hat{\bar{y}}$ is "design-consistent", so that the distribution, over repeated samples, of $\hat{\bar{y}}$ becomes increasingly concentrated around the true value $\bar{y}$ as the sample size grows larger.

The second assumption we make in using the interval (2.1) concerns the sampling design: It must be a *"probability sampling design"*, that is, one under which every unit in the finite population has a known positive probability, π_k, of being chosen. (It is intuitively clear that severely biased estimates could arise otherwise, namely, if we try to estimate the mean of all population units by a sampling procedure that gives some of the units no chance of being chosen.) The confidence level attached to (2.1) will under these assumptions be roughly 95 % provided the population is not too irregular in shape and the sample not too small.

For example, suppose the sampling design is stratified random sampling. If an auxiliary variable x is available, a traditional estimator of $\bar{y}$ is the "separate ratio estimator",

$$\hat{\bar{y}} = \frac{1}{N} \sum_{h=1}^{H} t_{xh} \frac{\bar{y}_{s_h}}{\bar{x}_{s_h}} \qquad (2.2)$$

(The strata are labelled $h=1,..., H$; t_{xh} is the known total of the auxiliary variable x in the h:th stratum, and $\bar{y}_{s_h}$, $\bar{x}_{s_h}$ the respective means of y and x in the sample s_h drawn by simple random sampling from the h:th stratum). The statistician may not recall offhand the formula for the estimated variance, but a look in any of the standard sampling texts will indicate that an appropriate formula to use in the calculation of the interval (2.1) is

$$\hat{V}_p(\hat{\bar{y}}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S^2_{es_h} , \qquad (2.3)$$

where $W_h = N_h/N$, n_h is the number of units drawn by simple random sample from the N_h units in the h:th stratum, and $S^2_{es_h}$ is the variance in s_h of the residuals $e_k = y_k - (\bar{y}_{s_h} / \bar{x}_{s_h})\, x_k$.

For the statistician it is, of course, trivial to interpret the interval (2.1) in the appropriate probability terms, but the user of statistics (who is probably more interested in the point estimate $\hat{\bar{y}}$ than in the interval) is less keenly aware of the meaning of this interval. To illustrate, we can imagine the following conversation between the Statistician and his client (the User), an interested layman. The Statistician has carried out a survey for the User; we enter the scene at a point where the Statistician has just presented the results of the survey, including a confidence interval calculation for $\bar{y}$.

User: I am satisfied with the precision indicated by your interval. After all, our budget permitted only a limited sample size. Incidentally, I recall from my statistics background that when you talk about 95 % confidence, it somehow measures the chances that the interval covers the unknown population mean.

Statistician: You are right. The probability is 95 % that our selection mechanism yields a sample for which the interval covers $\bar{y}$. More precisely, suppose we drew not one but 10 000 samples from the finite population, replacing each sample and using each time the sampling design that we have agreed on. For each sample, imagine that we calculate the corresponding confidence interval by the formula (2.1). Then about 9 500 of the 10 000 intervals will contain $\bar{y}$.

User: I know, of course, that you statisticians always make statements with less than full certainty. I have no difficulty with this. After all, our sample is only a small part of the finite population. And it seems entirely natural to me to associate this lack of full certainty with the repeated sampling process you managed to describe in a few words.

3. Design-based versus model-based inference

We need not hesitate about the meaning of the terms design-based inference and model-based inference. Equivalent definitions have been given several times in the literature. According to Särndal (1978), in design-based inference "the source of randomness is the probability ascribed by the sampling design to the various subsets of the finite population." By contrast, model-based inference is derived by "looking at the values $y_1,\ldots,y_N$ associated with the N units of the population as the realized outcome of random variables $Y_1,\ldots,Y_N$ having a N-dimensional joint distribution ξ, where the superpopulation ξ is modeled to reflect the available background knowledge". Two different probability distributions are involved. It is convenient to call the former "the p-distribution", the latter "the ξ-distribution".

Smith (1978) makes essentially the same distinction (without bringing in the terms design-based and model-based): "For survey analysis we can distinguish two principal contenders. These are (i) inferences based on the p-distribution generated by the randomisation in the design and (ii) inferences based on the ξ-distribution, a hypothetical distribution of errors associated with a stochastic model which is assumed to underlie the data."

The distinction has found its way into the *Encyclopedia of Statistical Sciences*, perhaps a sign that the definition can now be accepted with some degree of unanimity. Under the entry "Inference, design-based vs. model-based" (Koch and Gillings (1983)) we read: "Design-based and model-based inference are alternative conceptual frameworks for addressing statistical questions from many types of investigations. These include: ... Sample surveys of randomly selected subjects ... For sample surveys, the probabilistic interpretation of design-based inferences such as confidence intervals is in reference to repeated selection from the finite population via the given design." By contrast, model-based inferences are obtained via "a superpopulation with assumptions characterizing the actual finite population as one realization; and so their probabilistic interpretation is in reference to repetitions of the nature of this postulated sampling process."

Other recent references that discuss the distinction are Little (1983), Sundberg (1983), and Hansen, Madow and Tepping (1983).

Instead of "design-based inference" some authors prefer the synonymous terms "randomization inference" and "probability sampling inference"; see, for example, Hansen, Madow and Tepping (1983). (These authors reserve the term "sample design" to describe what several recent writers call a strategy, that is, the combined choice of a sampling design (or plan) and an estimator.)

It also follows that a natural although seldom-used synonym for "model-based inference" is "superpopulation inference". In other words, the distinction that we are examining could also be expressed with clarity as "randomization inference" versus "superpopulation inference."

Cases arise where one or the other kind of inference is more appropriate or in fact the only choice. Some populations are too extensive or complex to be sampled with a probability sampling scheme. Then model-based inference is the only possibility. An example would be a study of the prevalence of a health condition in general, rather than at a fixed point in time. It is not an exaggeration to say that the settings in which both types of inference are applicable exhibit considerable harmony between them. In what follows we shall see examples of (relatively minor) conflicts arising when both are applicable, as well as instances where one of them is strongly preferred or the only possibility.

4. The model-based interval

We have seen that the p-distribution underlies the design-based interval (2.1). The ξ-distribution is used to construct the model-based interval as follows: with reference to the model, we can claim that the unknown difference between the estimated mean and the true mean, $\hat{\bar{y}} - \bar{y}$, is a random quantity enclosed between the limits

$$\pm 1.96 \sqrt{\hat{V}_\xi(\hat{\bar{y}} - \bar{y})}$$

with 95 % level of confidence. The resulting interval for $\bar{y}$ is

$$\hat{\bar{y}} \pm 1.96 \sqrt{\hat{V}_\xi(\hat{\bar{y}} - \bar{y})}\,, \qquad (4.1)$$

where $\hat{V}_\xi(\hat{\bar{y}} - \bar{y})$ is a "good" estimate (from the point of the model) of the model variance, $V_\xi(\hat{\bar{y}} - \bar{y})$, of $\hat{\bar{y}} - \bar{y}$. This type of variance refers to repeated generation of populations $y_1,\dots,y_N$ obeying the model, given that a fixed set s has been drawn as a sample and that, consequently, the complement set, $\bar{s} = U–s$, is the "non-sample".

(Proponents of model-based inference sometimes point out that it is of no interest to consider, as one does in design-based inference, the samples s that could have been obtained, but were in fact not obtained. The proponent of design-based can retort: Why should one, as is done in model-based inference, consider populations of size N other than the one and only population that is actually present, since all but the actual one are purely hypothetical? Both kinds of inference are frequentist. They refer to repetitions: repeated samples in one case and repeated populations in the other. Now, if we were to use Bayesian inference, we may do away with repetitions altogether. But this is a different story; one that is not considered in this paper.)

After a moment of further reflection, we realize that (4.1) also requires that $\hat{\bar{y}} - \bar{y}$ is of zero (or essentially zero) expected value under the model; if not, the interval would systematically fail to contain $\bar{y}$.

In the example involving stratified sampling and the separate ratio estimator (2.2), suppose that the superpopulation model specifies that the y_k are independent random variables and that the regression of y_k on x_k passes through the origin with a separate slope in each stratum, so that

$$y_k = \beta_h x_k + \varepsilon_k$$

with

$$E_\xi(\varepsilon_k) = 0,\ V_\xi(\varepsilon_k) = \sigma_k^2 = x_k \sigma_h^2\,, \qquad (4.2)$$

for any unit k in stratum h. Here σ_h^2 is a "model variance" for stratum h. A calculation (of which we leave out the details) shows that

$$V_\xi(\hat{y} - y) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) a_h \sigma_h^2 \qquad (4.3)$$

with $a_h = \bar{x}_{U_h} \bar{x}_{\bar{s}_h} / \bar{x}_{s_h}$. (Here $\bar{x}_{U_h}$, $\bar{x}_{s_h}$ and $\bar{x}_{s_h}$ denote the respective means of x in the stratum, U_h, in the sample from the stratum, s_h. and in the non-sampled part of the stratum, $\bar{s}_h = U_h - s_h$.)

The modelist will now probably proceed by replacing the unknown σ_h^2 by its unbiased (under the model) estimate

$$\sigma_h^2 = \sum_{k \in s_h} \frac{(y_k - b_h x_k)^2}{x_k} / (n_h - 1)$$

where $b_h = \bar{y}_{s_h} / \bar{x}_{s_h}$. In his calculation of the model-based interval (4.1), he would thus use

$$\hat{V}_\xi(\hat{\bar{y}} - \bar{y}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) a_h \sigma_h^2 \qquad (4.4)$$

By contrast, the design-based interval was calculated according to (2.1) with the variance estimate (2.3).

In this example, both procedures use the same *point estimate*, but the *inferences* still differ: At a common 95 % level of confidence, the width of the two intervals will not be the same, since (2.3) differs from (4.4).

The difference can be numerically important. Which is the correct 95 % interval? Both are, of course, mathematically correct. The difference is caused by differing "bases of inference". (We note that model-based inference "depends" on the model in a very crucial manner; Hansen, Madow and Tepping (1983) therefore prefer the term "model-dependent inference".)

Let us return to the Statistician and the User. We now enter the scene at a later point where the statistician has just explained that, in addition to the design-based interval using (2.1) and (2.3), a model-based interval using (4.1) and (4.4) was also calculated, "just to see how they compare". Perhaps it would have been preferable not to mention this.

User: It seems to me that you are now involving two different notions of probability. It was all right with me to interpret the probability associated with the confidence interval as "the coverage rate in repeated samples". Frankly, what I mainly want is one number, the point estimate, $\hat{\bar{y}}$. In my work, I need the point estimate, and compared to this, the interval is of secondary interest. Sure, I want the estimate $\hat{\bar{y}}$ to be correct as far as possible, but how you methodologists arrive at guaranteeing the near-correctness is not my major concern. If both of your intervals for $\bar{y}$ are equally valid, I will naturally opt for the shorter one, if the need arises for an indication of precision. What you are saying about the two procedures is intriguing and confusing; unfortunately, I don't have time to sit down and really think through various interpretations of elusive concepts such as "probability" and "degree of confidence". The repeated sampling interpretation that you first gave seemed very palatable, although I can also follow the main lines of your second argument.

Statistician: You happened to strike on a very important word, namely, "valid". The model-based interval is valid only if the assumptions of the model are met. If the assumptions seem reasonable to you, we can settle for the model-based interval; if not, the design-based interval should be chosen, since it is valid regardless of any assumptions. Let me emphasize again that the two intervals represent two different conclusions about one and the same parameter, the mean $\bar{y}$ of the "real" finite population. I mention this because I am sure that in your work you sometimes feel the need to extend the conclusions beyond the finite population to the conceptual superpopulation expressed by a model. Then a model-based inference is natural.

Let us leave our two friends here; the Statistician has probably added further to the consternation of the User through his final comment implying that the User should specify if he wants conclusions about the actual finite population or about a superpopulation.

5. The role of assumptions

The *Encyclopedia of Statistical Sciences* goes on to point out that "design-based inferences involve substantially weaker assumptions than do model-based inferences". This interesting – and justified – statement deserves a comment, since the statistician is naturally inclined to work under the weakest possible assumptions. We have seen that it is the concept of variance that creates the principal difference between the two kinds of inference. Design-based inference refers to "variance under the p-distribution"; model-based inference to "variance under the ξ-distribution".

The two distributions are very different in nature: The p-distribution expresses the randomization rule actually used in the selection of a sample. By contrast, the ξ-distribution does not specify a rule of action, nor is it "used" to generate the finite population. It is only *assumed* that nature created the finite population according to the ξ-distribution. That is, the ξ-distribution is hypothetical, a model of the finite population. In their discussion of model assumptions, Hansen, Madow and Tepping (1983, p.790) say "... to ignore the type of inference that can be made without such assumptions takes unnecessary risks and may result in misleading inferences". The inference that does not take "unnecessary risks" is, obviously, the design-based one. Design-based inference is, if not valid regardless of any assumptions, at least valid under very weak assumptions. (We have pointed to the fact that the interval (2.1) requires, in order to be valid at the 95 % confidence level, that the population is not too extreme, etc.)

When an inference depends on assumptions, statisticians worry about the *robustness* of the conclusions, that is, the sensitivity of the conclusions to changes in the assumptions. Obviously, in model-based inference, robustness becomes a question of first-order importance. By contrast, in design-based inference, robustness is so to speak built into the procedure, and is therefore less of an issue.

We have seen that the design-based inference builds on only one kind of repetition, namely, repeated draws of samples s, under the given sampling design, from the fixed population. Design-based inference is *model-free*. (We have a parallel in another branch of statistics: so-called non-parametric inference can be said to be distribution-free.)

Model-based inference, too, needs only one kind of repetition, namely, repeated realizations of population vectors $(y_1,\ldots,y_N)$ according to the specifications of the model. There is no need to consider repeated draws of samples; the sample is fixed. As Smith (1981, p. 269) puts it: "In inference with respect to the ξ-distribution, the p-distribution does not enter." The model-based inference can be described as *design-free*.

Whereas the model-free property is seen by many as a strong selling point in favour of design-based inference, no one looking for "weaker assumptions" will be particularly encouraged to hear that model-based inference is design-free. The reason should be obvious, in light of the discussion earlier in this section: model-free means "assumption-free", whereas the design-free feature does not rid us of any assumptions. On the other hand, the design-free property can be seen as an advantage when conclusions are sought from samples selected by non-probability sampling schemes. An example is quota sampling. But in the absence of randomized selection, a number of delicate issues arise. Prudent sampling statisticians regard quota sampling with a skeptical eye.

Given then that "design-based inferences involve substantially weaker assumptions", the innocent reader may wonder why one would ever even discuss the possibility of model-based alternatives, with their "unnecessary risks" and possibilities of "misleading results".

It is beyond our scope to go into the detailed reasons why model-based inference is still considered. Suffice it to say that model-based inference has some strong points in its favour. Part of the reason is mathematical tractability: model-based inferences are often simple to develop, whereas in design-based inferences, the estimators are often biased in small samples and only asymptotically unbiased, etc. And, as already noted, the model-based variety of inference is just about the only possibility for non-probability samples.

Another reason is that design-based inference cannot always be relied on without supplementing the argument, at some point, with assumptions. A striking example occurs when the sample is reduced due to a certain rate of non-response. The non-response occurs according to an unknown distribution, and assumptions simply must be made about its nature. The design-based inference in this case becomes mixed with some model assumptions, a realistic attitude. A statistician who refuses to admit anything but purely design-based inference, if one can be found, can be said to believe that "deliberate randomization creates the only probability distribution appropriate for statistical inference". These are the words used by Royall (1983) to define what he calls the "Randomization Principle". To stick unequivocally to this principle is often impossible, as we have just remarked. Hansen, Madow and Tepping (1983), and many with them, favour design-based inference as far as it can be reasonably applied. Over a long history of survey sampling, this conduct has led to successful results. In my opinion, this attitude is quite different from that of a dogmatic believer in the Randomization Principle; I am not sure such a statistician exists. Royall (1983), on the other hand, denies design-based inference. The ideological debate is, however, beyond the scope of this paper; the interested reader is referred to Hansen, Madow and Tepping (1983), and to the discussion following the article. (In their purest forms, design-based inference and model-based inference are, as we have just seen, model-free and design-free, respectively. As Smith (1981) points out, there are also "hybrid theories" that combine elements of both kinds of inference. For example, Hartley and Sielken (1975) discuss a set-up for inference that involves a two-step sampling procedure: Step 1: Draw a "large sample" of size N from an infinite super population; Step 2: Draw a sample of size $n<N$ from the large sample of size N obtained in Step 1.)

6. Design-based approach, model-based approach: What is implied?

The terms design-based approach, model-based approach are often used in oral and written communication, but in a rather informal way. There does not seem to be a generally agreed-upon definition of the two terms (and perhaps none is necessary, if we accept informality). Nevertheless, let us examine the two terms, assuming that the design-based approach uses design-based inference and, correspondingly, that the model-based approach uses model-based inference. This is highly appropriate because after all, the way in which we make conclusions about the population is the cornerstone in any statistical approach. In the survey sampling context, the statistician's total approach contains a number of decision points. Reducing an *approach* to its bare essentials, let us say that it consists of three steps:

(a) the *choice of a sampling design* (or, as some say, a sampling plan), and the execution of this design to produce one single sample, s;

(b) the *choice of an estimator* (that is, is a mathematical formula thought to produce numbers that are, on the average, near the truth), and the actual calculation, by entering the sample data into the formula, of one single estimate $\hat{y}$;

(c) the *making of an inference* about one or more parameters of the finite population. This includes the choice of a variance estimator formula, the calculation of the variance estimate, and, finally, the calculation of a confidence interval or a coefficient of variation for the estimate $\hat{y}$.

Here, (a) and (b) together are often called a (sampling) *strategy*.

Let us first examine how the *design-based approach* handles the three steps (a), (b) and (c). We have argued that this approach uses design-based inference, which presupposes three elements mentioned in Section 2: (1) on the part of the sampling design that $\pi_k > 0$ for all k (so that we have a probability sampling design); (2) on the part of the estimator that it is design consistent; (3) on the part of the variance estimator that it refers to variability over repeated samples s, in other words, that variability is measured with reference to the p-distribution.

As long as these three elements are present in step (c), the design-based approach may take every possible advantage in steps (a) and (b) of relationships existing between the variable(s) of interest and auxiliary variables. This leaves considerable room for the statistician to use ingenuity and good judgment to model relationships and put them to work within the first two steps of the design-based approach.

If we assume that there is no non-response, can the design-based approach be model-free? That is, can the approach be carried out without relying at all on modeled relationships? Technically speaking the answer is yes, for none of the steps (a), (b) and (c) require a modeling effort. But practically speaking the answer is no: in a situation where some knowledge about relationships is available, it would be wasteful not to exploit such a possibility for improvement. A design-based approach with no element of modeling seems to result in only the most naive of procedures: a simple random sampling design, and estimation through the sample mean.

The following remark by Hansen, Madow and Tepping (1983, p. 778) is clearly concerned with steps (a) and (b) of the design-based approach: "... models of the population may suggest useful procedures for selecting the sample or the estimators. This is often done in probability sampling to great advantage". (For "probability sampling", we may substitute "design-based approach".) Kalton (1983) points out that "current practice makes considerable use of theoretical models to help in the design of samples, but seldom to the extent that the validity of the results depends on the models used". According to these points of view, models can be most helpful in steps (a) and (b), but to involve a model in step (c) may lead to invalid conclusions. Thus models guide the sampling strategy but their role ends there. If a model is reasonably correct, a good strategy will result.

Concerning step (a), the use of probability-proportional-to-size selection, say, at the first stage of a two- or multistage design, can be motivated by the assumption that the unknown cluster totals are roughly proportional to the size-measures; Horvitz-Thompson type estimation will then lead to a small variance. Other uses of model assumptions at the design stage include assumptions to guide the choice of cluster size in cluster sampling, or assumptions made in attempts to approximate the theoretically optimal Neyman allocation of a stratified sample.

A characteristic of an estimator considered suitable for the design-based approach is that it take into account the sampling design by attaching to each observation the appropriate "sampling weight". For the k:th unit, this weight is $1/\pi_k$, the inverse of the inclusion probability. If model assumptions were used to determine the sampling design, the estimator is hence "automatically" influenced by that model. This is true for the basic Horvitz-Thompson estimator,

$$\hat{\bar{y}} = N^{-1} \sum_{k \in s} y_k/\pi_k$$

as well as for simple or multiple regression estimators (see Section 8),

$$\hat{\bar{y}}=N^{-1}\{\sum_{k\epsilon s} y_k/\pi_k + \sum_{j=1}^{q} \hat{b}_j (\sum_{k=1}^{N} x_{jk} - \sum_{k\epsilon s} x_{jk}/\pi_k)\}. \tag{6.1}$$

In the latter case the relationship between y and $x_1,\ldots, x_j,\ldots,x_q$ enters explicitly into the formula.

As Hansen, Madow and Tepping (1983) put it, "... if we know enough about a relationship in the population we should use that knowledge. We can use it in ways such that the validity of the inference does not depend on the validity of any assumption". The estimator (6.1) appeals directly to the regression relationship, and is still appropriate for design-based inference. Special cases of (6.1) include ratio, simple and multiple regression estimators; the estimator (6.1) can thus be said to be determined by a design-model pair, see Section 8. Traditional sampling texts justify the classical ratio and regression estimators on correlational grounds only, with implicit rather than explicit modeling.

Let us now see how the *model-based approach* handles steps (a), (b) and (c). The inference will be model-based. As we observed in Section 4, this requires: (1) on the part of the estimator that it be consistent or unbiased under the model; (2) on the part of the variance estimator that it refers to variability over repeated realizations $(y_1,\ldots,y_N)$ under the model. An advocate of the model-based approach may also fix certain minimal requirements for the sample selection procedure, as a means of protection against model breakdown. Hansen, Madow and Tepping (1983) point out that the model-based approach (which they would probably rather describe as model-dependent) includes "a sampling plan that may or may not require randomization in sample selection", while "the estimators need not be randomization-consistent"; however, the approach "may have substantial advantages if the model is appropriate ... it may then be possible unequivocally to adopt a best sampling plan or best estimator."

The sample selection may thus be randomized in the model-based approach, but non-randomized sampling schemes are not excluded. Non-randomized schemes include judgment (quota) sampling, balanced sampling and purposive sampling. Such plans can be accommodated in the model-based approach, since the probabilistic structure derives from the model alone. However, they do not fit into the design-based approach, where the argument requires that known, positive inclusion probabilities be attached to the various units.

Can the model-based approach be entirely free of randomization elements? Theoretically speaking, this is possible. However, if the sample selection is not randomized, the modelist will take other measures to protect against model misspecification. One such procedure is balanced sampling, that is, the sample is chosen purposely to strongly resemble the population on key characteristics. One may require that the sample means of known control variables agree with their counterparts for the whole population.

We can summarize the two approaches as follows:

Approach	Sample selection	Choice of estimator	Inference (confidence interval)
Model-based	Randomized or non-randomized (e.g., balanced) sampling.	Optimal estimator, given the model.	Based on assumed ξ-distribution, thus dependent on its validity.
Design-based	Model may help; randomization is necessary.	Often guided by model to profit from auxiliary information.	Based on p-distribution; model-free.

7. Design-based estimator, model-based estimator: What is implied?

The terms design-based and model-based are abundantly used in the literature to qualify the word "estimator". In my opinion, it is in this connection that considerable confusion is likely to occur. What exactly is a design-based estimator, a model-based estimator? The literature does not converge on one widely accepted meaning of the two terms, just as was the case with design-based approach and model-based approach.

The most immediate interpretation is the following: a design-based estimator is one that is used, or can be used, in the design-based approach. Correspondingly, a model-based estimator is one that fits into the model-based approach. As we have argued, this requires certain basic properties of the estimator. The design-based estimator must be design-consistent or design-unbiased, and will therefore be expressed in terms of sample-weighted observations. The model-based estimator must be model-consistent or model-unbiased and will probably be optimal under the given model; it will most likely ignore the sampling weights and will therefore be biased under repeated sampling.

For example, Kalton (1983) certainly has in mind the design-based approach when he writes: "The large sample sizes typical of most surveys is the other characteristic that generally favors design-based estimators. Sample sizes are usually chosen to be large enough to provide estimators of sufficient precision for the main survey objectives ..." (Here "precision" obviously means "variance over repeated samples".) Kalton goes on to say: "Besides making model checking a major undertaking, the multipurpose and multivariate nature of surveys causes another difficulty for the model-based approach: the model-based estimator may be suitable for some statistics but not for others." It is implied that the model-based approach uses a model-based estimator.

Reasonable though they may seem, these interpretations of "design-based estimator" and "model-based estimator" are not beyond reproach.

A first difficulty is the obvious one that the same estimator can, in a given situation, be recommended by both approaches. Should such an estimator then be called design-based or model-based?

For example, consider the separate ratio estimator (2.2) discussed in Section 1. This classic estimator agrees with the design-based approach under the designs of simple random sampling and stratified simple random sampling; it also agrees with the model-based approach, under the model (4.2) which postulates a regression through the origin separately for each stratum. Thus the separate ratio estimator fits into both approaches: it is a design-based estimator for certain designs, and at the same time a model-based estimator for a certain model formulation.

There is another difficulty associated with the suggested interpretation. Estimators destined for the design-based approach can be derived by a formal technique (see Section 8) that builds directly on a model; the separate ratio estimator (2.2) is an example. The result is a design-based model-based estimator. Gibberish, some would say. Not at all, if we pay attention to each word. An estimator can perfectly well be *both* design-based *and* model-based, the former in the sense of fitting into the design-based approach, the latter in the sense of being constructed by means of a model.

"Design-based model-based estimator" sounds nonsensical because instinctively we interpret the design-based feature as being in opposition to, or excluding, the model-based feature, and vice versa. In Section 5, we found this exclusiveness to hold for the dichotomy "design-based inference" and "model-based inference": the former is model-free, the latter design-free. With respect to "estimator", however, the exclusiveness does not apply: the estimator in the design-based approach will often have model features. (It is less likely that the estimator in the model-based approach will have design features, such as sample-weighted observations.)

How do we eliminate the confusing term "design-based model-based estimator"? One possibility is to say "design-based estimator generated by a model"; then design-based estimator continues to mean one that is compatible with the design-based approach. Another possibility is to say "model-based randomization theory estimator", as do Hansen, Madow and Tepping (1983); then model-based signifies "generated by a model". One may also use the term "design-model estimator", as in Särndal (1981). This emphasizes that the estimator is determined by a design-model pair; Little (1983) stresses the need to consider survey inference as the study of design-model pairs.

8. Estimators formed from predicted values

Let us examine how estimators are constructed in the design-based approach and in the model-based approach. Little (1983) asks the following question (where a "D-modeler" is someone who restricts the use of models to the design stage): "How does the D-modeler estimate population quantities without modeling? I would say by the Horvitz-Thompson (HT) estimator and ratios thereof. Although regression and ratio estimates are included in most survey statisticians' armories, it is difficult to motivate them without reference to an underlying model ... The HT estimator, on the other hand, embodies the spirit of the design-based approach, which treats estimation as a problem of weighting rather than of prediction." Let the population total

$$t = \sum_{k=1}^{N} y_k \quad .$$

be the unknown characteristic to be estimated. The "weighting principle" is indeed epitomized by the Horvitz-Thompson estimator

$$\hat{t} = \sum_{k \epsilon s} y_k/\pi_k \qquad (8.1)$$

and the ratio variety thereof,

$$\hat{t} = N(\sum_{k \epsilon s} y_k/\pi_k) / (\sum_{k \epsilon s} 1/\pi_k) . \qquad (8.2)$$

In the vocabulary of some survey statisticians the term "design-based estimator" seems to be limited to these simple weighting estimators. However, to use modeling only in determining the design, and thereby the sampling weights $1/\pi_k$, is an unnecessarily restricted and inflexible use of relationships. As Hansen, Madow and Tepping (1983, p. 778) point out, in the design-based approach "models of the population may suggest useful procedures for selecting the sample *or* the estimators." (The emphasis is mine.)

Let us examine a formal procedure for building design-based estimators other than the basic weighting procedures (8.1) and (8.2). Predicted values play an important role in the following simple argument.

Let $x_1,\ldots,x_q$ be auxiliary variables such that (a) we have knowledge about $x_1,\ldots,x_q$that extends beyond the sampled units, for example, in the form of known totals

$$t_{xj} = \sum_{k=1}^{N} x_{jk} \quad ,$$

and (b) we can safely assume a strong regression relationship in the population, so that y is explained well by $x_1,\ldots,x_q$. If both (a) and (b) are present, we can use the x-variables to advantage. The values y_k are observed for the sampled units only.

An estimator for the design-based approach can be built as follows. Pretend for a moment that the regression coefficients $B_1,\ldots,B_q$ are known in the population. If the regression is strong,

$$y_k^0 = \sum_{j=1}^{q} B_j x_{jk}$$

will be a close approximation to y_k for most units k. The numerically small residuals

$$E_k = y_k - y_k^0$$

can then be obtained for units k in the sample.

Rewrite the population total to be estimated as

$$t = \sum_{k=1}^{N} y_k = \sum_{k=1}^{N} y_k^0 + \sum_{k=1}^{N} E_k ,$$

that is, as the sum of a known constant,

$$\sum_{k=1}^{N} y_k^0 = \sum_{j=1}^{q} B_j t_{xj} ,$$

and an unknown total of residuals, $\Sigma_1^N E_k$. An obvious estimator for the latter is $\Sigma_s\ E_k/\pi_k$, so a new improved estimator of t is given by

$$\hat{t}_{\text{DIF}} = \sum_{k=1}^{N} y_k^0 + \sum_{k \epsilon s} E_k/\pi_k .$$

The variance of $\hat{t}_{\text{DIF}}$ equals the variance of the sample sum of the weighted residuals, $\Sigma_s E_k/\pi_k$. Since the residuals are small, this variance is ordinarily much reduced compared to the variance of the sample sum of the weighted raw scores, $\Sigma_s y_k/\pi_k$. In other words, the variance of $\hat{t}_{\text{DIF}}$ is usually much smaller than that of the HT estimator.

In practice, $B_1,\ldots,B_q$ are unknown and replaced by estimates $b_1,\ldots,b_q$ based on the current sample. Let

$$(b_1,\ldots,b_q)' = \Big(\sum_{k \epsilon s} \underset{\sim}{x}_k \underset{\sim}{x}_k'/\sigma_k^2\pi_k \Big)^{-1} \sum_{k \epsilon s} \underset{\sim}{x}_k y_k/\sigma_k^2\pi_k ,$$

where σ_k^2 is the model variance of y_k and $\underset{\sim}{x}_k = (x_{1k},\ldots,x_{jk},\ldots,x_{qk})'$ is the value, for the k:th unit, of the vector of auxiliary variables. Then we can form the predicted values

$$\hat{y}_k = \sum_{j=1}^{q} b_j x_{jk}$$

and, for $k \epsilon s$, the residuals

$$e_k = y_k - \hat{y}_k .$$

The estimator becomes

$$\hat{t}_{\text{REG}} = \sum_{k \epsilon U} \hat{y}_k + \sum_{k \epsilon s} e_k/\pi_k .$$

This is the general form of a (multiple) regression estimator in the design-based approach; equivalently it may be expressed in the form (6.1).

Let us compare the approach above with the standard argument for building an estimator for the model-based approach. The total to be estimated is in this case rewritten as

$$t = \sum_{k=1}^{N} y_k = \sum_{k \epsilon s} y_k + \sum_{k \epsilon \bar{s}} y_k$$

where $\sum_{k \epsilon s} y_k$ is the known, observed sample total of y, and $\sum_{k \epsilon s} y_k$ is the unknown total for units k in the non-sampled part of the population, $\bar{s} = U\text{-}s$. We must estimate the latter sum. It is natural to replace each unknown y_k by the best possible "imputed value", $\hat{y}_k$, that we can construct, given $\underset{\sim}{x}_k$. The estimator for the model-based approach then becomes

$$\hat{t}_{\text{MOD}} = \sum_{k \epsilon s} y_k + \sum_{k \epsilon s} \hat{y}_k ,$$

where

$$\hat{y}_k = \sum_{j=1}^{q} \hat{\beta}_j x_{jk}$$

and

$$(\hat{\beta}_1,\ldots,\hat{\beta}_q)' = \Big(\sum_{k \epsilon s} \underset{\sim}{x}_k \underset{\sim}{x}_k'/\sigma_k^2 \Big)^{-1} \sum_{k \epsilon s} \underset{\sim}{x}_k y_k/\sigma_k^2$$

is the vector of estimated regression coefficients. Note that the term $\sum_{k \epsilon s} \hat{y}_k$ can be calculated from our auxiliary information, since

$$\sum_{k\in\bar{s}} \hat{y}_k = \sum_{j=1}^{q} \hat{\beta}_j(t_{xj} - \sum_{k\in s} x_{jk}),$$

where the totals $t_{x1},\ldots,t_{xq}$ are known, and $\hat{\beta}_j$ and $\sum_{k\in s} x_{jk}$ are calculated from the sample for $j=1,\ldots,q$. For example, under the model (4.2) both $\hat{t}_{\text{REG}}$ and $\hat{t}_{\text{MOD}}$ give the separate ratio estimator

$$\hat{t}_{\text{REG}} = \hat{t}_{\text{MOD}} = \frac{1}{N}\sum_{h=1}^{H} t_{xh}\frac{\bar{y}_{s_h}}{\bar{x}_{s_h}} \qquad (8.3)$$

provided units are sampled with equal probability within each stratum, so that $\pi_k = n_h/N_h$ for all units k in stratum h. If not, $\hat{t}_{\text{REG}}$ and $\hat{t}_{\text{MOD}}$ will differ. And, as we noted in Section 4, the confidence intervals that accompany the estimator (8.3) are not identical in the design-based and model-based approaches.

We do not imply that $\hat{t}_{\text{REG}}$ and $\hat{t}_{\text{MOD}}$ will always agree. In more complex situations, the two formulas often exhibit differences, as in estimation for domains; see Särndal (1984).

9. Conclusion

We have argued that statistical terms arising from the distinction design-based versus model-based are not always clear. An exception is "design-based inference" and "model-based inference"; for these terms generally accepted definitions exist. However, when the distinction is applied to other concepts, such as "approach" or "estimator", it is not automatically clear what is meant. The paper can be seen as a step in the direction of clarification of terminology, as well as a warning against uncritical use of terms that invoke the distinction.

10. References

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983): An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. (With discussion.) Journal of the American Statistical Association, 78, pp. 776–807.

Hartley, H.O. and Sielken, R.L., Jr. (1975): A "Super-Population Viewpoint" for Finite Population Sampling. Biometrics, 31, pp. 411–422.

Holt, D., Smith, T.M.F. and Winter, P.D. (1980): Regression Analysis of Data from Complex Surveys. Journal of the Royal Statistical Society, Series A, 143, pp. 474–487.

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

Koch, G.G. and Gillings, D.B. (1983): Inference, Design Based vs. Model Based. In S. Kotz, N.L. Johnson and C.B. Read (editors): Encyclopedia of Statistical Sciences, Vol. 4. New York: Wiley, pp. 84–88.

Little, R.J.A. (1982): Models for Nonresponse in Sample Surveys. Journal of the American Statistical Association, 77, pp. 237–250.

Little, R.J.A. (1983): Comment on M.H. Hansen, W.G. Madow and B.J. Tepping: An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. Journal of the American Statistical Association, 78, pp. 797–799.

Rennermalm, B. (1980): Biasskattning för regressionskoefficienter i design-baserad inferens från ändliga populationer. Statistisk tidskrift, No. 2, pp. 126–128. (In Swedish.)

Royall, R.M. (1983): Comment on M.H. Hansen, W.G. Madow and B.J. Tepping: An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. Journal of the American Statistical Association, 78, pp. 794–796.

Särndal, C.E. (1978): Design-Based and Model-Based Inference in Survey Sampling. (With discussion.) Scandinavian Journal of Statistics, 5, pp. 27–52.

Särndal, C.E. (1980): On π-Inverse Weighting versus Best Linear Weighting in Probability Sampling. Biometrika, 67, pp. 639–650.

Särndal, C.E. (1981): Frameworks for Inference in Survey Sampling with Applications to Small Area Estimation and Adjustment for Nonresponse. Bulletin of the International Statistical Institute, 49:1, pp. 494–513.

Särndal, C.E. (1982): Implications of Survey Design for Generalized Regression Estimation of Linear Functions. Journal of Statistical Planning and Inference, 7, pp. 155–170.

Särndal, C.E. (1984): Design-Consistent versus Model-Dependent Estimation for Small Domains. Journal of the American Statistical Association, 79, pp. 624–631.

Särndal, C.E. and Wright, R.L. (1984): Cosmetic Form of Estimators in Survey Sampling. Scandinavian Journal of Statistics, 11, pp. 146–156.

Smith, T.M.F. (1978): A Model Building Approach to Survey Analysis. Paper presented at the European Meeting of Statisticians, Oslo, August, 1978.

Smith, T.M.F. (1981): Regression Analysis for Complex Surveys. In D. Krewski, R. Platek and J.N.K. Rao (editors): Current Topics in Survey Sampling. New York: Academic Press, pp. 267–292.

Sundberg, R. (1983): The Predictive Approach and Randomized Population Type Models for Finite Population Inference from Two-Stage Samples. Scandinavian Journal of Statistics, 10, pp. 223–238.

Wright, R.L. (1983): Finite Population Sampling with Multivariate Auxiliary Information. Journal of the American Statistical Association, 78, pp. 879–884.

Received September 1984
Revised January 1985