A prediction approach to representative sampling

Ib Thomsen and Li-Chun Zhang[1]

Abstract

After a discussion on the historic evolvement of the concept of representative sampling in official statistics, we propose a prediction approach to it.

1 The birth of representative method

The word birth should not be taken literally, as there are many fathers and few (if any) mothers involved in the process. As a matter of fact, social and demographic statistics emerged as a result of partial investigations, rather than census-like investigations. In Stephan (1947) is given a number of early statistical analyses based on incomplete investigations. In Kruskal and Mosteller (1979a,b,c) they investigate the origin of "representative investigations", and give several examples, some of which go as far back as "times of Athens". Finally, in Seneta (1985) is elucidated the origins and development of the representative method in the Russian Empire/Soviet Union.

The authors do not agree on all issues, but there seem to be a general agreement that the Norwegian statistician A. N. Kiær may be credited with the real "fatherhood" of the use of the representative method in statistical practice. The Russian statistician concludes in his book (Kaufman, 1913): "The honour- if not of the discovery- but in any case of the systematic treatment of the method, is essentially due to the Norwegian statistician Kiær who in the nineties of the last century made this method the basis for a number of inquiries and acted as a zealous propagandist of the same ..."

Kiær introduced the method for an international audience at the ISI meeting in Bern, August 1895. He presented results from a survey based on the 1890 Census. The design used was what to day would be called a three-stage design. In the first stage, 128 counties and 23 towns were selected throughout the country. In the second stage, a sample of males with ages 17, 22, 27, 32 etc. was selected. Finally, in the third stage, males who's name started with the letters A, B, C, L, M , N were selected. The 1890 Census was used as a frame. To evaluate the representativity of the sample, he produced a number of tables of marginals from the survey, and compared the results with the census tables.

In his conclusions from the meeting, he stated: "I find that 1. the sampling methods should be of great importance for the development of statistics chiefly in the special representative investigations are arranged so that they can be controlled with regard to the chief points with

[1]Statistics Norway, Kongensgate 6, PB 8131 Dep, N-0033 Oslo, Norway. E-mail: lcz@ssb.no

help of general investigations, 2. that according to the circumstances different methods might be concerned from among which one should have to choose and 3. that consequently the advantages and the difficulties of the different methods deserve to be recommended for being studied and discussed by the statisticians."

The reception of these ideas was generally negative at the meeting. Several of the well established statisticians of that time were negative, and the method was not recommended. Nevertheless Kiær achieved two important goals at the meeting:

- He made sure that discussions concerning the method should continue at the next ISI meeting, and thereby making it a focal forum for the development of the method.
- Through this meeting and his broad international network of colleagues, he inspired many statisticians throughout the world to use and to investigate the method.

During the following ISI meetings, Kiær continued to present results from representative surveys conducted in Norway. The real meaning of representativeness was still not clear. Kiær seemed to have in mind a sample as a proportionally scaled down miniature population, where each unit in the sample is close to the same number of units in the population. This can be seen from his emphasis on making marginal comparisons between the sample and the census.

In 1924, the ISI formed a committee to investigate the feasibility of the representative method. The Danish statistician Jensen, was rapporteur of this committee. He stated at the following ISI meetig: "When ISI discussed the matter twenty two years ago, it was the question of the recognition of the method in principle that claimed most interest. Now it is otherwise. I think I may venture to say that nowadays there is hardly one statistician, who in principle will contest the legitimacy of the representative method. Nevertheless, I believe that the representative method is capable of being used to a much greater extent than now is the case". In other words, the representative method was born and kicking!

It is worth mentioning that the method was accepted by official statisticians before the theory was developed. Only a few theoretical papers were published, and most of them were unknown to the committee. An exception is the paper by Bowley (1926). Bowley was member of the aforementioned committee, and probably knew about results he published in 1926.

Allow us a little detour at this point into the field of register based official statistics. The situation in this field today is similar to that of sampling in 1926. After a long period of scepticism among official statisticians, it is now well recognized that data collected for administrative purposes, can play an important role in production of official statistics. This approach is accepted in spite of the fact that few statistical concepts are developed in this area. Another similarity between the two fields is that the scepticism they both were received with by official statisticians was due to a threat to the budgets of the statistical institutions.

2 Rise and fall of the representative method: Balance vs. randomization

In Kruskal and Mosteller (1980) they traced the history of the concept of representative sampling in 1895 - 1939. In their summary of "the great debates" during the ISI meetings before 1925, they state: "To sum up, Kiær pressed his ideas at the Meetings in Bern, St. Petersburg, Budapest and Berlin and in other writings and he used the method. He had some opposition and some support. He sharpened his description during this period. Von Bortkiewicz introduced and applied a significance test for representativeness, but since Kiær had no probabilistic model, the test had to be 'as if' he had a specific one. March widened the discussion and developed the idea of probability sampling".

Kiær came to call his approach the 'representative method', and his 1895 paper is entitled 'Observations and experiences with representative surveys'. His intuition is perhaps more clearly revealed during the 1903 meeting where Kiær made an innovative remark: "The difference between the ratios in a representative sample and in the population are scarcely larger than those between the results from one full census to another". This idea about variability over time of the population itself, i.e. to view the population as arising from an underlying stochastic process, fits of course under the so-called 'model-based' approach. Kiær did not take a probability sampling point of view. For him it was not so much about representative sampling but representative sample. The closeness in the marginal comparisons between the sample and the census for a number of control variables, which he so much stressed, might in modern terminology be called 'multivariate simple balance'. To aim at such a balanced sample is entirely justifiable provided (i) the population can be divided into a moderate number of homogenous sub-populations, and (ii) the same variation exists in all these sub-groups. In other words, a group mean model with a constant variance across the population groups.

The subsequent development, however, first took a turn away from Kiær's approach. The papers by Bowley (1926) and in particular by Neyman (1934) are usually taken as the starting point of the theoretical development for the so-called 'design-based' approach. A number of earlier papers do however exist, notably by the Danish statistician Gram (1883) and the Russian statistician Chuprov (1923). In both papers, the optimal allocation of observations among strata was given many years before Neyman's paper. These papers were unfamiliar to a wide audience, probably because they were written in Danish and Russian, respectively. In any case, representative sampling became inseparable with randomization as Neyman states,

> Thus, if we are interested in a collective character X of a population π and use methods of sampling and of esimation, allowing us to ascribe to every possible sample, Σ, a confidence interval $X_1(\Sigma)$, $X_2(\Sigma)$ such that the frequency of errors in

the statements

$$X_1(\Sigma) \leq X \leq X_2(\Sigma)$$

does not exceed the limit $1 - \epsilon$ prescribed in advance, *whatever the unknown properties of the population*, I should call the method of sampling representative and the method of estimation consistent.

In other words, 'no randomization, no representativeness'. Equal probability selection is no longer a necessity, but probability sampling is.

After the Neyman paper, the theory of probability sampling developed fast. Pioneers like Hansen, Hurwitz and Madow combined theoretical and practical work from the late 30's. They published their results in a number of articles, and finally collected their findings in their famous three volumes textbook. Other textbooks followed, notably Deming, Kish and Cochran. Among developing countries, Mahalanobis and Lahiri (1961) used sampling in the field of agricultural statistics. The earlier discussions quoted above concerned essentially with social statistics. In other areas as business statistics and agricultural statistics a more general approach was needed. The reason is that the population units are more heterogeneous there than in social statistics. Furthermore, one had access to some kind of measure of size for each unit before selection.

Within the design-based framework, Godambe (1955) showed that no minimum variance estimator of the total exists within the general class of linear estimator, where the coefficients in front of the observations may vary from one unit to another as well as from one sample to another. This negative result seem to have caught the attention of many theoreticians, and a number of papers on the foundation of sampling appeared. Many of these contributions are collected or referred to in Johnson and Smith (1969).

In many of these papers, a model was introduced for the population, under which the study variable y in the population is a stochastic variable, and the population total T is therefore also a random variable. Let t be an estimator of T, then the properties of the error $(t - T)$ may be studied under repeated sampling as well as under the model for y. To illustrate this approach, we shall introduce a simple model, which we shall use during the rest of this presentation. Let the population consist of N units, with labels $1, 2, ..., N$. To each unit is associated two values (x, y), where x is known for all units in the population. The aim is to estimate

$$T = \sum_{i=1}^{N} y_i$$

To estimate T a sample s of size n is selected using a randomization design. We assume that the y-values are realized values of independent, random variables $Y_1, Y_2, ..., Y_N$, such that

$$Y_i = \beta x_i + \epsilon_i \tag{1}$$

where β is an unknown coefficient and the ϵ_i's are independent random variables with

$$E(\epsilon_i) = 0 \qquad \text{and} \qquad E(\epsilon_i^2) = \sigma_i^2 = \sigma^2 x_i^\gamma \quad \text{for} \quad \gamma \geq 0$$

In this formulation, we have two random processes, one generated by the method of sample selection, and one by the model (1). A number of strategies were developed that combine a sampling method with an estimator. Interested readers are referred to Chapters 12 and 14 in Särndal, Swensson, and Wretman (1992) for an accessible account.

In 1970, Royall published a small, but important paper in which he introduced the prediction approach to sampling (Royall, 1970). Here is suggested to base the inference only on the model. Under this approach, the y-values outside the sample are predicted using e.g. the model (1). A predictor of the total, may now be written in the following way:

$$t = \sum_{i=1}^{n} y_i + \hat{\beta} X_r$$

where $\hat{\beta}$ is an estimate of β under model (1), and $1, ..., n$ are the sample units, and $X_r = \sum_{i=n+1}^{N} x_i$ is the total of x-values outside of the sample. This formulation of t is appealing as it uses the observed values of y, and predicts the unobserved. It leads to a number of papers on the foundation of sampling (e.g. Smith, 1976; Cassel, Särndal, and Wretman, 1983).

In our case, we are concerned with how to select the sample. It is clear that all non-informative samples produce valid estimates of $\hat{\beta}$ under the model alone and therefore valid predictors of the total T as well. A choice that minimizes the mean squared error (MSE) of t leads to a non-probabilistic or purposive sample of the n units with the largest x-values in the population, for all $0 \leq \gamma \leq 2$. As the sample depends on the model, it is vulnerable to model failures. In Royall and Eberhardt (1975) is suggested to balance the sample to make the model-based prediction more robust, i.e. remains unbiased despite certain mis-specifications of the model. This however does not require that the sample is selected with known probability. For instance, provided $\bar{x} = \bar{X}$, where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{X} = \sum_{i=1}^{N} x_i/N$, the simple expansion estimator $t = \sum_{i=1}^{n} y_i N/n$ remains unbiased under the model (1) above, and the ratio estimator $t = (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{N} X_i)/(\sum_{i=1}^{n} x_i)$ under the model (1) remains unbiased even if the true population model contains an intercept term in the linear predictor, and so on.

Such samples of simple balance seem to be just what Kiær was aiming at by his representative method, because simple balance is achieved in any sample that is truly a miniature population, where each sample unit corresponds to the same number of units in the population. Kiær used the sample means as estimators of the population means, and looked for multivariate simple balance, i.e. multiple x-values instead of a single one. This has the same effect as using the expansion estimator based on a sample of simple balance, which remains unbiased for any

y-variable that can be related to any of the x-values by model (1) above.

Finally, the theory of simple balance applies when the population units have a constant variance. Situations where the individual variance $V(Y_i|x_i)$ varies across the units call for the theory of weighted balance. However, this requires adopting models that are not as intuitive as the ratio model (1) above. The practitioners have so far seemed reluctant to do so.

3 A prediction approach to representative sampling

In what follows we shall outline a prediction approach to representative sampling. Under this approach we can in principle predict each unit in the population, and we suggest that representative sampling is more readily connected to the *individual* MSE of prediction (IMSEP), which is given by

$$\text{IMSEP}_k = E_p\{E[(\hat{Y}_k - Y_k)^2|s]\} \tag{2}$$

where $\hat{Y}_k$ is a predictor of Y_k. For $k \in s$, we have $\hat{Y}_k = Y_k = y_k$ and there is no error of prediction; whereas the conditional MSE of prediction is positive for all units outside of the sample. Thus, the unconditional IMSEP is generally positive for all the population units except those with inclusion probability one. In this way, a randomization design can be used to control the unconditional prediction accuracy of each unit in the population, which can be regarded as the *expected* amount of information we may obtain for it under the assumed model.

Let us look at an example. Let the population consist of two units. Each unit is associated with a variable of interest Y_i and an auxiliary variable x_i. Assume the model (1) above. Let the sample size be $n = 1$. Suppose the desired *control of individual prediction (CIP)* criterion is given as

$$\text{IMSEP}_k \propto \lambda_k \qquad \text{for} \quad k = 1, 2$$

for some chosen constants λ_k. Given the sample, we have $\hat{\beta} = Y_i/x_i$ for $i \in s$. If $i = 1$, then we have $\hat{Y}_2 = (Y_1/x_1)x_2$ and $\hat{Y}_2 - Y_2 = \epsilon_1(x_2/x_1) - \epsilon_2$, such that

$$\Delta_2 = E[(\hat{Y}_2 - Y_2)^2|i = 1] = \sigma^2(\frac{x_2^2}{x_1^2}x_1^\gamma + x_2^\gamma) \quad \text{and} \quad \text{IMSEP}_2 = E_p(\Delta_2) = \pi_1 \cdot \Delta_2$$

since $E[(\hat{Y}_2 - Y_2)^2|i = 2] = 0$ if $i = 2$. Similarly, if $i = 2$ we have

$$\Delta_1 = E[(\hat{Y}_1 - Y_1)^2|i = 2] = \sigma^2(\frac{x_1^2}{x_2^2}x_2^\gamma + x_1^\gamma) \quad \text{and} \quad \text{IMSEP}_1 = E_p(\Delta_1) = (1 - \pi_1) \cdot \Delta_1$$

Sampling design under three different CIP criteria and assuming the model (1) with $\gamma = 0, 1, 2$ are given below in terms of the π_i:

	CIP criterion		
Model (1)	$\lambda_1 = \lambda_2$	$\lambda_1/x_1 = \lambda_2/x_2$	$\lambda_1/x_1^2 = \lambda_2/x_2^2$
$\gamma = 0$	$x_i^2/(x_1^2 + x_2^2)$	$x_i/(x_1 + x_2)$	1/2
$\gamma = 1$	$x_i^2/(x_1^2 + x_2^2)$	$x_i/(x_1 + x_2)$	1/2
$\gamma = 2$	$x_i^2/(x_1^2 + x_2^2)$	$x_i/(x_1 + x_2)$	1/2

Notice that the last two of these are, respectively, sampling with probability proportional-to-size and equal probability sampling. In other words, these familiar design can be motivated from a prediction point of view under a chosen CIP criterion and an assumed model.

It is seen that the proposed prediction approach to representative sampling leads to randomization designs. It should, however, be pointed out that randomization does not target directly at efficiency of estimation. For instance, if the aim is to estimate the total of the y-values, the most efficient choice of sample would be to select the unit with the largest value of x_1 and x_2. Neyman perceived randomization as providing safeguard against bias in sample selection. Varying the probability of selection may reduce the sampling variance, as in the case of stratification with dis-proportionate allocation of stratum sample sizes. But there is no such thing as a minimum variance sampling design in the general sense, as Godambe's result suggests.

Zhang and Thomsen (2007) develop prediction designs that balance between the control of individual prediction and efficient prediction of the population totals under the general linear model. The aim is not any single 'optimal' design *per se*. Rather, the balancing approach generates sets of nested designs that form a basis on which reasonable choices can be made in practice. Moreover, it provides a unified framework which leads naturally to a number of well established, but seemingly unconnected sampling techniques in practice.

- Under a homogeneity model with a constant mean and a constant variance for all the units in the population, equal individual prediction requires the use of equal probability selection method, such as simple random sampling (SRS).

- Stratified SRS is generally a stratified equal prediction design under a stratified homogeneity model. Relative equal individual prediction, i.e. $\text{IMSEP}_k/V(Y_k) \propto 1$, implies proportional allocation of stratum sample size, regardless of the stratum population variance and/or the variable of interest.

- Under the ratio model (1) commonly used in business surveys, the prediction approach leads to the division of take-all, take-some and take-none units. The take-alls are the units with the largest x-values, created to secure a baseline efficiency for the population totals. The take-nones are the units with the smallest x-values, created because the intrinsic prediction uncertainty, i.e. $\sigma_k^2 = V(\epsilon_k)$, is so small for these units that the corresponding IMSEP may be smaller than that of a remaining take-some unit despite the take-nones

have no chance of being selected. Finally, stratified relative equal prediction criterion leads to stratified SRS among the take-some units.

- Under the common parameters model with a constant population mean, variance and intra-cluster correlation, it can be shown that the two basic two-stage sampling schemes, namely PPS-SRS and SRS-SRS designs, are equal prediction designs provided the intra-cluster correlation is 0 or 1, respectively. Stratified SRS-SRS two-stage sampling, with first-stage size stratification of the primary sampling units, provides means for balancing between prediction for total and individuals.

We think there are three principal advantages in introducing the control of individual prediction as a design criterion.

- It establishes the model-based approach as a legitimate mode of inference. Individual prediction is impossible under the design-based framework of inference, because there is no model that connects the observations to the unobserved values. The only exception is the trivial case where the sample contains all the units in the population (or a sub-population) apart from a single unit, in which case a prediction can be made by subtracting the observed total from the estimated population total. This is a major inadequacy of the design-based framework of inference.

- It initiates the need for randomization design, which is an essential aspect of the concept of representative sampling. Equal probability sampling, for instance, is not necessary for model-based prediction of the population total, even under the population common mean model. It can prevent selection bias, but so can any non-informative sampling design. It does aim at a sample of simple balance, but it is not the most effective device for that. But equal probability sampling is needed if we would like to have the same IMSEP for all the population units under the population common mean model.

- In practice, for effective use of available resources, a sampler frequently has to balance among several conflicting aims of a sampling design. Balancing between the prediction of the totals and the individuals of a given population offers a systematic approach to this. While the former aims at the most aggregated level of data, the latter aims at the most disaggregated level. The need for good population totals is of course a starting point for most of the surveys conducted at the national statistical offices. Moreover, for a long time now, appropriate micro-data have been vital for social-economic research and planning, such as in micro-simulation modeling (Orcutt, 1957). The need can not be satisfied by census data alone. Finally, at an intermediate level, small area (or domain) estimation (Rao, 2003) has received a lot of attention in the past decades, reflecting the growing demand of such statistics for fund allocation, regional planning, and business decision making. Due to the large number of small areas it is usually impractical to treat them as design strata. Also, users may ask for multiply defined domains of interest, which overlap with each other and can not all be planned in advance. But it is possible to derive and

control the properties of a sampling design for small domain estimation through the prediction of individuals (Zhang and Thomsen, 2005), because these are the smallest possible domains.

References

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, **22**.

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys*, vol. **3**, pp. 143–160. New York: Academic Press.

Chuprov, A.A. (1923). On the mathematical expectation of the moments of frequency distribution in the case of correlated observations. *Metron*, **2 (3)**, 461–493 (646–683).

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, **17**, 269–278.

Gram, J.P. (1883). Om beregning af en bevoxnings masse ved hjelp af prvetrær (In Danish). *Tidsskrift Skovbruk*, **6**, 137–198.

Johnson, N.L. and Smith, H. (1969). *New Developments in Survey Sampling*. New York; Wiley.

Kaufman, A. (1913). Theorie und metoden der Statistikk. Mohr, Tubringen.

Kruskal, W. and Mosteller, F. (1979a). Representative sampling, I: Nonscientific literature. *International Statistical Review*, **47**, 13–24.

Kruskal, W. and Mosteller, F. (1979b). Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review*, **47**, 113–127.

Kruskal, W. and Mosteller, F. (1979c). Representative sampling, III: the current statistical literature. *International Statistical Review*, **47**, 245–265.

Kruskal, W. and Mosteller, F. (1980). Representative sampling, IV: the history of the concept in statistics, 1895 - 1939. *International Statistical Review*, **48**, 169–195.

Mahalanobis, P.C. and Lahiri, D.B. (1961). Analysis of errors in censuses and surveys with special reference to experience in India. *Bulletin of the International Statistical Institute*, **38**.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558–606.

Orcutt, G.H. (1957). A new type of social-economic system. *Review of Economics and Statistics*, **39**, 116–123.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Seneta, E. (1985). A sketch of the history of survey sampling in Russia. *Journal of the Royal Statistical Society, Series A*, **148**, 118–125.

Smith, T.M.F. (1976). The foundations of survey sampling. *Journal of the Royal Statistical Society, Series A*, **139**, 183–204.

Stephan, F.F. (1947). History of the uses of modern sampling procedures. In *Proceedings of the ISI Conference*, vol. III of *Part A*. International Statistical Institute.

Zhang, L.-C. and Thomsen, I. (2005). A prediction view on sampling design: Clustered population and two stage-sampling. *Invited talk at SAE2005: Challanges in Statistics Production for Domains and Small Areas.*

Zhang, L.-C. and Thomsen, I. (2007). A prediction approach to sampling design. In preparation.