Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

USING AUXILIARY INFORMATION TO CHOOSE BETWEEN ALTERNATIVE SAMPLING DESIGNS IN A SURVEY WITH SEVERAL KEY VARIABLES

Anders Holmberg[1]

ABSTRACT

At the design and estimation stages of a survey, large survey organizations often use auxiliary information. Technological advances in data capture and a better accessibility of registers open up for an increased and more efficient use of such information. This paper addresses issues of how to use auxiliary information efficiently in sampling from finite populations. Since different auxiliary variables may have varying strength for different target parameters, and since a design that is the best for one parameter not necessarily is the best for another, good compromise designs are needed.

In this paper, previous results regarding the choice of optimal design are extended to the case of several study variables. By minimizing different summary measures at the planning stages, we suggest an approach to achieve a good compromise design that has a high overall efficiency. Precision requirements that are different for different estimates are handled by applying a non-linear programming algorithm. Our approach yields a set of unequal first order inclusion probabilities, which can be applied with both fixed size and random size sampling schemes and it offers a flexible use of auxiliary variables in the design. By way of an example based on an application to a Swedish business population we compare the approach with single-variable routines and other methods used by practising survey statisticians to address the multivariate problem in the design.

KEY WORDS: Auxiliary Variables; Multivariate Surveys; Optimal Design; Survey Planning.

1. INTRODUCTION

1.1 Preliminaries

The typical concern of a survey is the estimation of a large number of parameters. One important goal for the survey statistician in the planning phase of the survey is to find a strategy – a combination of sampling design and estimators for the most important parameters – which gives the best possible (most precise) parameter estimates for a given cost, or conversely to obtain estimates with a desired precision at lowest possible cost. Survey sampling theory offers instruments to achieve such a goal and in this paper we will illustrate and compare methods that survey statisticians may use at the planning stage to handle a multivariate situation.

In the search for estimators of high precision, survey sampling theory strongly emphasizes the use of auxiliary information. If, as frequently is the case in large survey organizations, there is auxiliary information available, the statistician can use this information to his advantage and thereby obtain a highly efficient strategy. Often, the sampling frame is equipped with at least a few auxiliary variables that may be useful, while subject matter experts and the survey statistician may have knowledge of the relationship between study variables and auxiliary variables from similar surveys of similar populations. If no such useful knowledge is available, it will often be worthwhile to make a pilot survey. In this paper, 'auxiliary information' will be reserved for the following specific kind of information, consisting of two related components: (i) data on a set of variables (which will be called auxiliary variables) available at the planning phase as well as at the estimation phase for every population element; (ii) appropriate a priori knowledge of the structure of relationships between important study variables and (subsets of) the auxiliary variables.

[1] Anders Holmberg, Statistics Sweden, Klostergatan 23, SE-70189 Örebro, Sweden, anders.holmberg@scb.se

Because auxiliary variables can be used in a wide variety of ways at the design stage as well as at the estimation stage, efforts to use this information to find efficient estimators lead to many techniques for sampling and estimation. (Results on optimum allocation in stratified sampling, on probability proportional-to-size sampling, and on regression or calibration estimators all rely on properties of auxiliary variables.) However, whereas the auxiliary information for the estimators may be individually selected and different for different estimators, the auxiliary information used in the design is shared, and will affect all the parameter estimates. Hence, when a multivariate survey is planned, the design choice is – since it affects all estimates – relatively more important than the choice of estimators. Although a good estimator sometimes can compensate for a poorly chosen design and give an acceptable precision, it is also likely that precision is lost compared to what is achievable with the same estimator and a better design. Before implementing a design that uses auxiliary information, the statistician must therefore closely examine its effects on all the key estimators. This paper focuses on examining such effects and on how to find a good compromise when there are many ways in which the auxiliary information may be used. The theoretical content of the paper is a summary of previous results on optimal designs and here they are applied on a Swedish business population.

1.2 The Problem and some Notation

Initially, some explanations of the theoretical background and the notation are needed. Assume that we are planning a survey of a finite population $U=\{1,\ldots,k,\ldots,N\}$, and assume that the key parameters we want to estimate are the population totals of the unknown study variables $y_1,\ldots,y_q,\ldots,y_Q$ i.e. $\mathbf{t}=\left(t_{y_1},\ldots,t_{y_q},\ldots,t_{y_Q}\right)'$, where $t_{y_q}=\sum_{k\in U} y_{qk}=\sum_U y_{qk}$. A without replacement sampling design $p(\bullet)$, with first-order inclusion probabilities π_k $\left(k=1,\ldots,N\right)$ and second-order inclusion probabilities π_{kl} $\left(k,l=1,\ldots,N\right)$, will be used to select a random set sample $s\subseteq U$ of size n_s and each of the population totals are to be estimated with estimators $\hat{\mathbf{t}}=\left(\hat{t}_{y_1},\ldots,\hat{t}_{y_q},\ldots,\hat{t}_{y_Q}\right)'$.

One of the criteria when the sampling design is selected should be a precision criterion. Normally, such a criterion is defined by functions of the estimator variances. For example, the design is to be determined to minimize a function f of all Q estimator variances and fulfill specified restrictions, v_q, made on functions g of each estimator variance (or variance approximation), i.e. minimize,

$$f\left(g(V(\hat{t}_{y_1})),g(V(\hat{t}_{y_2})),\ldots,g(V(\hat{t}_{y_Q}))\right) \tag{1}$$

with restrictions on

$$g\left(V\left(\hat{t}_{y_q}\right)\right)\le v_q\left(q=1,\ldots,Q\right) \tag{2}$$

To be of practical use in survey planning, a general criterion as the one above needs to be simplified, more concrete and adapted to what is known and available, essentially some auxiliary information. Here, it is assumed that there are P auxiliary variables accessible at the planning stage. They are denoted $u_1,\ldots,u_p,\ldots,u_P$ and their values u_{pk} $\left(p=1,\ldots,P\right)$ are known for every element *k* in the population.

Another aid in the planning process comes from using statistical models. Hence, if useful a priori knowledge about the relations between the study variables and the auxiliary variables exists, then models which describe those relations well are worked out. An often fruitful example is to formulate linear models, ξ_q, $\left(y_{qk}=\mathbf{x}'_{qk}\boldsymbol{\beta}_q+\varepsilon_{qk}\right)$ for the study variables, with $E_{\xi_q}\left(\varepsilon_{qk}\right)=0$, $V_{\xi_q}\left(\varepsilon_{qk}\right)=\sigma^2_{qk}$ and $E_{\xi_q}\left(\varepsilon_{qk}\varepsilon_{ql}\right)=0$ $\left(k\ne l\right)$, i.e.,

$$\begin{aligned} E_{\xi_q}\left(y_{qk}\right) &= \mathbf{x}'_{qk}\boldsymbol{\beta}_q \\ V_{\xi_q}\left(y_{qk}\right) &= \sigma^2_{qk} \end{aligned} \tag{3}$$

where $\mathbf{x}'_{qk} = \left(x_{1_q k}, \ldots, x_{j_q k}, \ldots, x_{J_q k}\right)$ is a suitable set of J_q (positive) auxiliary variables formed from $u_1, \ldots, u_p, \ldots, u_P$, and where $\boldsymbol{\beta}_q = \left(\beta_{1_q}, \ldots, \beta_{j_q}, \ldots, \beta_{J_q}\right)'$ and σ^2_{qk} are model parameters.

The models are helpful since – when we plan the survey to achieve a specified precision criterion – we need a substitute for the unknown estimator variances. Moreover, we need some measure that can help us to discriminate and choose between alternative strategies. This measure must be a function of auxiliary information, since that is the only information available. The anticipated variance, defined by Isaki (1970) (see also Isaki and Fuller (1982)) is one such measure. If $\hat{t}_{y_q}$ is an estimator of t_{y_q} and if (3) is interpreted as moments of a superpopulation from which a finite population has been selected, the anticipated variance is the variance of $\hat{t}_{y_q} - t_{y_q}$ over both the model ξ_q and the sampling design, i.e.,

$$E_{\xi_q} E_p\left[\left(\hat{t}_{y_q} - t_{y_q}\right)^2\right] - \left[E_{\xi_q} E_p\left(\hat{t}_{y_q} - t_{y_q}\right)\right]^2 .$$

The anticipated variance might be used to compare the properties of different strategies. In this paper, the point estimators $\hat{\mathbf{t}}$ are members of the GREG (Generalized regression) family. The GREG estimator is defined as

$$\hat{t}_{y_q r} = \hat{t}_{y_q \pi} + \left(\mathbf{t}_{x_q} - \hat{\mathbf{t}}_{x_q \pi}\right)' \hat{\mathbf{B}}_q . \tag{4}$$

Here, $\hat{t}_{y_q \pi} = \sum_{k \in s} y_{qk} / \pi_k = \sum_s y_{qk} / \pi_k$ is the well-known Horvitz-Thompson or π estimator, $\mathbf{t}_{x_q} = \left(t_{x_{1q}}, \ldots, t_{x_{jq}}, \ldots, t_{x_{Jq}}\right)'$ is a J_q-dimensional vector of x_q totals, $\hat{\mathbf{t}}_{x_q \pi}$ is a vector of the corresponding π estimators and

$$\hat{\mathbf{B}}_q = \left(\sum_s \frac{\mathbf{x}_{qk}\mathbf{x}'_{qk}}{c_{qk}\pi_k}\right)^{-1} \sum_s \frac{\mathbf{x}_{qk} y_{qk}}{c_{qk}\pi_k} \tag{5}$$

is an estimated vector of regression coefficients, where c_{qk} is a suitable constant. (Details of GREG estimation are given in Särndal, Swensson and Wretman (1992) sections 6.4-6.7.)

2. PLANNING FOR AN OPTIMAL SAMPLING DESIGN

2.1 Some Theoretical Background in the Univariate Case

If a model ξ_q is well specified, then an approximation to the anticipated variance of $\hat{t}_{y_q r}$ can be written as $ANV_q(\hat{t}_{y_q r}) = \sum_U \left(\pi_k^{-1} - 1\right)\sigma^2_{qk}$. The design that minimizes $ANV_q(\hat{t}_{y_q r})$ (for a given q , a given expected sample size and under the restrictions $0 < \pi_k \le 1$) is a design where $\pi_k \propto \sigma_{qk}$ $\left(k = 1, \ldots, N\right)$. For fixed size designs, other authors have presented results also indicating that an optimal design is attained when $\pi_k \propto \sigma_{qk}$ – Hajek (1959) (linear design-unbiased estimators), Brewer (1963) (ratio estimation) and Cassel, Särndal and Wretman (1976) (generalized difference estimators). When the π estimator coincides with the ratio estimator Godambe (1955) also shows a related result.

In most presentations, the concept of 'optimum' is tied to a single study variable. Examples are to be found in the ($\pi_k \propto \sigma_{qk}$)-theory above (normally leading to a choice of unequal inclusion probabilities), in the theory of optimum stratification and optimum allocation in stratified sampling and in the theory of multi-stage sampling, where the number of units in each stage can be determined optimally. (Reviews of various results on optimality with concern to survey sampling designs are given by Rao (1979) and Bellhouse (1981).) However, it is not straightforward how to extend these examples to the practical planning of a multivariate situation with several study variables with varying relations to several auxiliary variables. In the following we suggest one approach.

2.2 An Extension to the Multivariate Case

A simple reason why few papers deal with the multivariate survey problem might be that there is no evident criterion of optimality. An example on how we may apply the ($\pi_k \propto \sigma_{qk}$)-theories in the planning of a multivariate case is the following. First we reformulate our criterion (1) and substitute the unknown estimator variances by the approximated anticipated variances $ANV_q(\hat{t}_{y_q r})$ ($q = 1,\ldots,Q$) and since we only have planning stage means, we need to replace the unknown σ_{qk}^2 by guesses (guesstimates) of the variance structure. We will denote guesstimates used for planning purposes with the tilde symbol e.g. $\tilde{\sigma}_{qk}^2$. (Brewer (2002) chapter 7 discusses commonly occurring and reasonable choices of $\tilde{\sigma}_{qk}^2$.) If we have good guesstimates and apply the ($\pi_k \propto \sigma_{qk}$)-theories; a preferred sampling design from a univariate perspective, would be a design characterized by a set of preferred first-order inclusion probabilities where $\pi_k \propto \tilde{\sigma}_{qk}$. However, in the multivariate case there might be one such preferred set for every key variable, and perhaps there are as many different 'optimal' designs to choose from, as there are key variables. In such a situation a careful investigation of the overall effect of different design alternatives should be done before the final design decision is taken. In many cases we must (due to some kind of multivariate consideration) find a compromise design.

A reasonable step in an investigation of various design effects is to compute, for every q $\left(q = 1,\ldots,Q\right)$, the set of preferred first-order inclusion probabilities needed to minimize $ANV_q(\hat{t}_{y_q r})$. The values in these sets will be denoted by $\tilde{\pi}_{q(opt)k}$. Hence, they are planning phase calculations based on guesstimates and the assumed model relationships. The minimum of $ANV_q(\hat{t}_{y_q r})$ to be used at the planning phase is then given by $ANV_{q\min}(\hat{t}_{y_q r}) = \sum_U \left(\tilde{\pi}_{q(opt)k}^{-1} - 1\right)\tilde{\sigma}_{qk}^2$ for $\left(q = 1,\ldots,Q\right)$.

As a multivariate optimisation criterion the objective function to minimize, f, could simply (as below) be a (possibly weighted) arithmetic mean of relative ratios that considers the precision of all the $\hat{t}_{y_q r}$. This is then a kind of loss function which we call the Anticipated Overall Relative Efficiency Loss (*ANOREL*), and it is defined as

$$ANOREL = \sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_q r})_{p_i}}{ANV_{q\min}(\hat{t}_{y_q r})}. \quad (6)$$

Where H_q $(q = 1,\ldots,Q)$ are weights (summing up to unity) that reflect the relative importance of the parameters to be estimated. $ANV_q(\hat{t}_{y_q r})_{p_i}$ is the approximation to the anticipated variance of $\hat{t}_{y_q r}$ under a design $p_i(\bullet)$ with $\pi_k = \pi_{p_i k}$. The minimum of *ANOREL* for a given sample size is obtained if

$$\pi_k \propto \sqrt{\sum_{q=1}^{Q} H_q \frac{\tilde{\sigma}_{qk}^2}{\sum_U \left(\tilde{\pi}_{q(opt)k}^{-1} - 1\right)\tilde{\sigma}_{qk}^2}} \tag{7}$$

However, if restrictions, v_q, (upper limits) of the ratios are imposed i.e.

$$\frac{ANV_q\left(\hat{t}_{y_q r}\right)_{p_i}}{ANV_{q\,\min}\left(\hat{t}_{y_q r}\right)} \le v_q \quad q = 1,\ldots,Q \tag{8}$$

then the minimisation of (6) becomes a non-linear optimisation problem. This non-linear optimisation problem to be solved can be written as in formulas (9) and (10): Minimise the objective function,

$$f(\boldsymbol{\pi}) = \sum_{q=1}^{Q} H_q \sum_U (\pi_k^{-1} - 1) \frac{\tilde{\sigma}_{qk}^2}{\sum_U (\tilde{\pi}_{q(opt)k}^{-1} - 1)\tilde{\sigma}_{qk}^2} \tag{9}$$

subject to the $2N + Q + 1$ restrictions

$$\begin{aligned} & 0 < \pi_k \le 1 \quad k = 1,\ldots,N \\ g_0(\boldsymbol{\pi}) &= \sum_U \pi_k - n = 0 \\ g_q(\boldsymbol{\pi}) &= \sum_U \left(\pi_k^{-1} - 1\right) \frac{\tilde{\sigma}_{qk}^2}{\sum_U \left(\tilde{\pi}_{q(opt)k}^{-1} - 1\right)\tilde{\sigma}_{qk}^2} \le v_q \quad q = 1,\ldots,Q \end{aligned} \tag{10}$$

Examples of other objective functions and a detailed description of the optimisation model are found in Holmberg (2002) and Holmberg, Flisberg and Rönnqvist (2003).

Remark: The planning approach to come up with a good strategy for a sample survey adopted here results in a decision on a class of *preferred sampling designs* characterized by a preferred set of first-order inclusion probabilities. To implement a design conforming to this set we need a *sample selection scheme*. We will not discuss this issue, but there are various solutions to the question, e.g. random size schemes such as Poisson sampling or if fixed size schemes are preferred then the Rao-Sampford scheme or Pareto sampling (independently proposed by Saavedra (1995) and Rosén (1997)) are alternatives. If (at least approximately) unbiased variance estimation not is considered an important issue then systematic sampling is another alternative.

2.3 Comments on Methods which Deal with the Design Choice in the Multivariate Case

What are then the advantages of using this multivariate extension outlined above? It seems as if there are three often used ways to approach the multivariate problem in practice: (i) Ignore it and reduce it to a single variable problem through a snap judgment based on ‘experience’. (ii) For each of the important parameters, study the effect of various single-variable dependent actions, and then select a design that seems to be the best compromise. (iii) Find some kind of overall multivariate criterion, perhaps mechanically by generalizing single variable concepts (as in section 2.2 here), and choose the action that optimises this criterion.

The first choice may not be as bad as it may sound. In many situations, a few important study variables may have similar properties, and the total loss in precision following from focusing on one of them might be acceptable. However, the second approach is better since it means that a design choice is explicitly made on the basis of considerations of the multivariate situation. An example that fits into this approach is a method described by Kott and Bailey (2000) called *maximal Brewer selection*. It is not an optimisation-based method, but it is simple to implement and it guarantees the desired precision of all estimators considered in the planning. Various methods of averaging key measures may also be placed in this category of approaches. One is to select a stratified sample with an allocation based on the average of various single-variable optimum allocations. Another, in the case of probability proportional-to-size sampling with σ_q as size measure, is to choose the size measure as the average or median of Q different single variable σ_q’s.

If we consider the third (optimisation-based) approach category a number of authors have presented solutions for multivariate stratified sampling, e.g. Dalenius (1957), Chatterjee (1968), Hughes and Rao (1979), Chromy (1987). Other relevant references concerning the third approach are: Sigman and Monsour (1995) who sketched a procedure using non-linear programming similar to that of section 2.2 for Poisson πps sampling and the π estimator, and Saavedra (1999) who applied these ideas using the algorithm proposed by Chromy to determine probabilities to be used for Pareto πps sampling in a price and volume petroleum product survey.

The advantage of using the third kind of approach is that, at least theoretically, optimal solutions are obtained. Moreover, as we choose the optimisation criterion, we also attain certain control of the desired design properties, as opposed to relying on intuitive ad hoc solutions. However, as for all other approaches the success will ultimately depend on the chosen criterion and the underlying assumptions made by the survey designer.

3. AN ILLUSTRATION OF THE MULTIVARIATE EXTENSION

3.1 A Description of the Studied Application

We illustrate the multivariate extension of the $\pi_k \propto \sigma_{qk}$ theories from section 2.2. with a simple application on a Swedish business population. The data is an extract from an administrative register of Swedish business enterprises and in the example below we study enterprises within the branch 'Manufacturers of metal goods (except machines and devices)'. For the elements in the population, $k = 1, \ldots, 2292$, we have the values of four auxiliary variables, *Number of Employees,* u_{1k}, *Turnover,* u_{2k}, *Personnel expenses,* u_{3k}, and *Investments,* u_{4k}. Suppose that the aim with the survey is to estimate the yearly population totals of these variables for a later reference time than that of the register data. Hence, except for the reference time, the variable definitions of the four auxiliary variables matches that of the $q = 1, \ldots, 4$ study variables so that t_{y_1} is the total number of employees, t_{y_2} is the total turnover and so on. Previous experience suggests that for Number of employees, Turnover and Personnel expenses, a simple ratio model fairly well describes the relations between the auxiliary variable and the study variable. Hence, for $q = 1, 2, 3$ our (univariate) planning stage models are

$$\begin{aligned} E_{\xi_q}(y_{qk}) &= \beta_q u_{qk} \\ V_{\xi_q}(y_{qk}) &= \sigma_{qk}^2 = u_{qk} \end{aligned} \tag{11}$$

Consequently we use u_{qk} as our guesstimates $\tilde{\sigma}_{qk}^2$, for $q = 1, 2, 3$. For the investment variable $q = 4$, scatterplots and estimations made on data from earlier years indicate that a heteroscedastic variance component such as the one in (11) is not motivated. As an alternative for the investment variable we use $\tilde{\sigma}_{4k}^2 = 1$. The four study variables are considered equally important and we suppose our budget allows a sample with an expected sample size of $E_p(n_s) = 290$. Moreover, we require that our sampling design is such that no $ANV_q(\hat{t}_{y_q r})_{p_i}$ exceeds $ANV_{q\min}(\hat{t}_{y_q r})$ by more that 10%, i.e. $v_q = 1.1 \ \ (q = 1, 2, 3, 4)$. By adding this last requirement, it is unlikely that any design other than a solution of the optimisation problem (9-10) is satisfactory, but we cannot be certain unless some planning phase comparisons are made.

3.2 Computation of Relative Efficiency Losses at the Planning Phase

In order to make a suitable design choice, we prepare some planning phase diagnostics that illustrate the properties of our alternative designs. One example of such diagnostics is to compute anticipated efficiency losses for the various designs under consideration. If we do this in our multivariate business survey application we get the following results.

Given the conditions stated in section 3.1, there are six alternative designs that are close at hand. Let $p_1,\ldots,p_4$ be the designs that follow if we choose $\pi_k \propto \tilde{\sigma}_{qk}$, i.e. the four designs we get by computing $\tilde{\pi}_{q(opt)k}$ $(q = 1,2,3,4)$. Let p_5 be the design following from choosing π_k according to equation (7) and let p_6 be a design resulting from solving the optimisation problem given by equations (9) and (10). The diagnostics of these designs are summarized in Table 1, where the cells contain the relative efficiency loss, i.e. $100[(ANV_q(\hat{t}_{y_q r})_{p_i} \big/ ANV_{q\min}(\hat{t}_{y_q r})) - 1]$.

Table 1: The Anticipated Relative Efficiency Losses of Six Considered Sampling Designs in percent.

	Study variables				
Considered Design	y_1	y_2	y_3	y_4	*Mean*
p_1	0	7.3	2.0	21.0	7.6
p_2	6.1	0	4.3	33.0	10.9
p_3	1.8	5.0	0	24.4	7.8
p_4	31.6	51.2	36.1	0	29.7
p_5	1.7	4.9	1.9	14.0	5.6
p_6	3.4	7.0	4.0	10.0	6.1

Not surprisingly, we note from table 1 that design p_1 (optimal for the Number of employees) and design p_3 (optimal for Personnel Cost) are very alike. They are expected to work well for the estimation of t_{y1} and t_{y3} and quite well for t_{y2}. Similar properties (but with a higher mean efficiency loss) are indicated by the anticipated efficiency losses of the p_2 design. None of these three designs are satisfactory for the investment variable y_4. The compromise designs p_5 and p_6, based on the multivariate extensions and optimising ANOREL in equation (6) are better in that respect. The mean efficiency loss of the p_6 design (6.1 %) is higher than that of p_5. That is the price we pay for satisfying the restriction that $ANV_q\left(\hat{t}_{y_q r}\right)_{p_1} / ANV_{q\min}\left(\hat{t}_{y_q r}\right) \leq 1.1$ for every $q = 1,2,3,4$.

4. FINAL REMARKS

It is only the multivariate optimisation-based design, p_6, that is satisfactory in the application above, and it has a better overall efficiency than the designs based on single-variable optimisation. To a practitioner, cumbersome computations are a possible drawback with an optimisation-based approach. Fortunately, with a computer program, the proposed multivariate extension is only slightly more complicated than single-variable methods. In addition, the proposed method is to a large degree based on computations that should be executed anyway in any serious planning effort. An important point of the multivariate extension is the flexible solution on how to use auxiliary information exhaustively in the design planning. The survey statistician is free to use any suitable combination of auxiliary information for all the key variables. However, when we face a multivariate survey problem, it has to be stressed that it is the work done in obtaining and analysing the diagnostics to support the final design choice that is the most important part. Statistics that illustrate design effects from a multivariate perspective are especially useful, not only when the survey statistician makes a design decision, but also in a discussion with non-statisticians taking part in such decision. With a non-linear optimisation program it is also both fast and relatively easy to make new computations if we for example want to investigate the effects of changes in the restrictions and sample size.

REFERENCES

Bellhouse, D. R., (1984). A review of optimal designs in survey sampling. Canadian Journal of Statistics, 12 pp 53-65.

Brewer, K.R.W. (1963). Ratio Estimation and Finite population: Some results deductible from the assumption of an underlying stochastic process. Australian Journal of Statistics 5, 93–105.

Brewer, K.R.W. (2002). Combined Survey Sampling Inference, Weighting Basu’s Elephants. Arnold, London.

Cassel, C.M., Särndal, C-E. and Wretman, J. (1976). Some results on generalized difference estimators and generalized regression estimators for finite populations. Biometrika 63, 615–620

Chatterjee, S. (1968). Multivariate stratified surveys. Journal of the American Statistical Association 63, 530–534.

Chromy, J. (1987). Design Optimization with Multiple Objectives. Proceedings of the Section on Survey Research Methods, American Statistical Association 1987 194–199.

Dalenius, T. (1957). Sampling in Sweden. Almquist & Wiksell, Stockholm

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

Hajek, J., (1959). Optimum strategy and other problems in probability sampling. Casopis pro Pestovani Matematiky 84 387–421.

Holmberg, A., (2002). A multiparameter perspective on the choice of sampling design in surveys. Statistics in Transition, 5, no 6, pp. 969–994.

Holmberg, A., Flisberg, P., and Rönnqvist, M., (2003). On the choice of optimal design in business surveys with several important study variables. In: Holmberg, A., (2003). Essays on Model Assisted Survey Planning. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 126, Doctoral thesis Uppsala University, Sweden.

Hughes, E., and Rao, J.N.K. (1979) Some problems of optimal allocation in sample surveys involving inequality constraints. Communications in Statistics A 8, 1551–1574.

Isaki, C.T. (1970). Survey designs utilizing prior information. Doctoral thesis, Iowa State University, Ames, Iowa.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model, Journal of the American Statistical Association, 77, 89–96

Kott, P.S. and Bailey, J.T. (2000). The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling, Proceedings of the second International Conference on Establishment Surveys, June 17–21, 2000, Buffalo 269–279.

Rao, J.N.K., (1979). Optimization in the design of sample surveys. In: J.S. Rustagi (ed.), Optimization methods in Statistics: Proceedings of an International Conference. New York, Academic Press, pp 419-434

Rosén, B., (1997). On sampling with Probability Proportional to Size, Journal of Statistical Planning and Inference, 62, 159-191.

Saavedra, P.J. (1995). Fixed Sample Size PPS Approximations with a Permanent Random Number. Proceedings of the section on Survey research Methods Joint Statistical Meetings, American Statistical Association, 697-700.

Saavedra, P.J. (1999). Application of the Chromy Algorithm with Pareto Sampling, Proceedings of the Section on Survey Research Methods Joint Statistical Meetings, American Statistical Association 1999 355–358.

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

Sigman, R.S. and Monsour, N.J. (1995). Selecting Samples from List Frames of Businesses, in Cox, B.G., Binder, D.A., Chinnappa, N., Christianson, A., Colledge, M.J., and Kott, P.S. (eds) Business Survey Methods, New York: Wiley, 153–169.