Statistics Canada Statistique Canada

# SURVEY METHODOLOGY

December 1981
Volume 7
Number 2

A Journal produced by
Methodology Staff
Statistics Canada

Canada

# SURVEY METHODOLOGY

December 1981 Vol. 7 No. 2

A Journal produced by Methodology Staff, Statistics Canada

## CONTENTS

8-3200-501
Reference No.
Z - 079

ISSN: 0714-0045

SURVEY METHODOLOGY

December 1981 Vol. 7 No. 2

A Journal produced by Methodology Staff, Statistics Canada.

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed; however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 6th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested.

# NOTES ON INFERENCE BASED ON DATA FROM COMPLEX SAMPLE DESIGNS

Gad Nathan[1]

The problems associated with making analytical inferences from data based on complex sample designs are reviewed. A basic issue is the definition of the parameter of interest and whether it is a superpopulation model parameter or a finite population parameter. General methods based on a generalized Wald Statistics and its modification or on modifications of classical test statistics are discussed. More detail is given on specific methods-on linear models and regression and on categorical data analysis.

## 1. INTRODUCTION

Standard methods of inference, such as regression, analysis of variance or tests of independence, are, in general, based on the assumption that the data are obtained by simple random sampling from an infinite population with a probability distribution belonging to some hypothetical family. The wide dissemination of standard computer packages has made the use of these methods extremely easy. However standard methods cannot usually be simply applied to data from complex sample designs without any modification.

In the following we attempt to provide a selection of some practical hints on what can be done and of some warnings against what should not be done in these situations. This is based on the selected list of references to recent work in the area, which include many examples of applications.

The first question which must be answered by anyone who intends to carry out statistical analysis is what exactly are the parameters about which inference is required.

[1] G. Nathan, Hebrew University, Jerusalem and Isreal Central Bureau of Statistics

One of two extreme answers to this question is often given (Brewer and Mellor (1973); Smith (1976)). One, as advanced for instance by Kish and Frankel (1974), considers that the only relevant inference concerns <u>finite population parameters</u>, such as the population regression coefficient:

$$B = \sum_{i=1}^{N} (X_i-\bar{X})(Y_i-\bar{Y}) / \sum_{i=1}^{N} (X_i-\bar{X})^2,$$

similarly defined multiple or partial correlation coefficients or other measures, defined with respect to the finite population only, with no recourse to any superpopulation model. Inference in this case would usually be design-based (Sarndal (1978)), that is based only on properties of the sample distribution. However model-based inference about a finite population parameter is also possible (Hartley and Sielken (1975)).

The other extreme position, as stated, for instance, by Fienberg (1980), considers all inference as relating to the <u>parameters of a probability distribution</u> (a superpopulation) of which the finite population represents a realization. Examples of such inference can be found in Konijn (1962), Fuller (1975), Thomsen (1978) and Pfeffermann and Nathan (1981). If the parameters about which inference is made relate to a superpopulation model, design-based inference cannot be used alone and inference must be model-based, Sarndal (1978), or jointly model- and design-based. Under assumptions of independence between the model distribution and the sampling distribution, standard (model-based) inference is valid and the sample design only affects the efficiency of inference.

Serious objections can be raised with respect to each of these extreme approaches. Model-based inference relies heavily on assumptions about a theoretical model which are usually difficult to ensure and the inference will not, in general, be robust to departures from this model. On the other hand, the finite population parameters, on which design-based inference is made, are usually "copies" of theoretical model parameters with little descriptive value in themselves, unless some basic model is assumed. For instance, a finite population correlation coefficient is a useful measure of the relationship between two variables only if the relationship is approximately linear.

In many cases some balance between these approaches may be preferable. This can be attained, for instance, by considering as the objects of inference only finite population parameters which closely approximate superpopulation parameters of a suitable model,to which the data fit. For instance, if separate regression equations are fitted to relevant sub-populations a better linear fit may be obtained than from an over-regression. If the sub-populations are large enough this will ensure that the finite population regression coefficients closely approximate the superpopulation parameters, so that any inference relating to the finite population parameters can be considered as relating to the superpopulation parameters.

To ensure close correspondence between model parameters and finite population parameters extensive exploratory analysis to check the model should be carried out,before entering into any formal analysis. This analysis to explore various alternative models can often be based on simple descriptive measures for which the sample design can be taken into account or on graphical displays. However the results have to be carefully interpreted in the light of the sample design. For example, a few large residuals with small sample weights may be much less important than many smaller residuals with large weights. A useful diagnostic tool to consider in the case of regression is the difference between a weighted and an unweighted regression coefficient. A large difference will often indicate that the model is inadequate.

Once the parameters have been determined,we should consider what type of inference is required (point estimation, interval inference or tests of hypotheses). While point estimation and confidence intervals would be most appropriate for finite population parameters,tests of hypotheses, and in particular simple hypotheses, are strictly relevant only with respect to superpopulation parameters of a well-defined model. For example the hypothesis that two domain means are equal can only be seriously entertained with respect to the superpopulation means rather than their finite population realizations. If one wishes to avoid the formulation of a model it would be preferable to use point estimation or confidence intervals for the difference between the domain means rather than tests of hypotheses. If hypothesis testing about finite population parameters is required,testing a composite hypothesis (e.g. that the difference between the means is in a given range of values) would be more appropriate than testing the simple hypothesis(that the difference is zero). Note that for sufficiently large samples, any non-zero difference, no matter how small, will be found significantly different from zero.

In the following, we discuss some basic general methods of analysis of data from complex sample designs and some specific methods for linear models and for tests of goodness of fit and of independence in contingency tables. In general we shall consider the inference as relating to finite population parameters. However we consider this inference as relevant only if the finite population parameters closely approximate superpopulation model parameters. This leaves open the possibilities of tending either towards a purely design-based approach or towards a purely model-based approach, according to one's personal degree of belief in the validity of an underlying model.

## 2. BASIC GENERAL METHODS

### 2.1 Generalized Wald Statistic

If the hypothesis to be tested is linear (or can be linearized) in the expected values of asymptotically normal statistics, for which a consistent estimator of the variance matrix is available, the generalized Wald Statistic can be used (Grizzle, Starmer and Koch (1969)), Koch, Freeman and Freeman (1976), Freeman, Freeman, Brock and Koch (1976), Shah, Holt and Folsom (1977) and Koch, Stokes and Brock (1980)).

We assume that we wish to test the hypothesis:

$$H_o: \quad X\beta = \theta_o, \qquad (2.1.1)$$

where X is a known r×p design matrix of full rank. β is a p×1 unknown parameter vector (either finite population parameters or superpopulation parameters) and $\theta_o$ is a known r×1 vector of constants. In case the hypothesis is not linear a first-order Taylor series approximation can be used (Nathan (1972) and Shuster and Downinq (1976)).

We assume that a consistent asymptotically normal estimator $\hat{\beta}$, of β is available, as well as a consistent estimator, $\hat{V}$, of the covariance matrix of $\hat{\beta}$, whose distribution is independent of that of $\hat{\beta}$.

Then the qeneralized Wald Statistic, defined as:

$$X_w^2 = (X\hat{\beta}-\theta_o)' \ (X \hat{V} X')^{-1} \ (X\hat{\beta}-\theta_o) \qquad (2.1.2)$$

is asymptotically distributed, under the null hypothesis, as chi-square with degrees of freedom equal to the dimension of the hypothesis (p-r).

The consistency of $\hat{\beta}$ and of $\hat{V}$ and the asymptotic distributions of $\hat{\beta}$ and of $X_w^2$ can all be considered with respect to the sampling distribution or with respect to the superpopulation distribution.

The major problem associated with this approach is in obtaining the consistent estimator, $\hat{V}$, of the covariance matrix when $\hat{\beta}$ is nonlinear in the sample observations (as will often be the case). Rao (1975) surveys the various methods of variance estimation which can be used: linearization (Tepping (1968)); Balanced Repeated Replication (McCarthy (1969)); and Jackknife (Miller (1974)). Several general computer programmes are available for their implementation - e.g. SUPERCARP (Hidiroglou, Fuller and Hickman (1980)), SUDAAN (Shah (1978)) for linearization and OSIRIS IV: PSALMS for balanced repeated replication. A complete listing and comparison of programs is given by Kaplan, Francis and Sedransk (1979).

Empirical comparisons of the variance estimators are given by Kish and Frankel (1974) and by Richards and Freeman (1980) and theoretical comparisons by Krewski and Rao (1981).

However, attention should be given to the stability of the variance estimator, especially when the number of parameters is large. In addition, care must be taken with respect to the conditions under which consistency and asymptotic properties hold for complex designs. For instance, for a two-stage design asymptotic results may require both a large number of PSU's and a large number of final units per PSU.

## 2.2 Approximation and Modelling of the Covariances

The practical difficulties involved in obtaining a stable consistent estimator of the covariance matrix have led to attempts to use simplified approximations to such estimators. The basic idea is that by assuming some structure for the covariance matrix, more stable estimators of fewer parameters can be used.

The approximation can be carried out under a pure design-based approach, directly with respect to the covariance matrix. If assumptions can be made on equality of design effects for variances and covariances within a given sub-group of parameters,overall estimators of covariance can be used. This approach is used, for instance, by Nathan (1973), Fuller and Rao (1978), Fellegi (1980) and Lepkowski and Landis (1980).

Alternatively modelling of the population structure itself can lead to simplified covariance matrices which can easily be estimated (see, e.g., Altham (1976), Fuller and Battese (1973), Tomberlin (1979), Holt, Richardson and Mitchell (1980), Imrey, Sobel and Francis (1980) and Pfeffermann and Nathan (1981)).

### 2.3 Modifications of Standard Tests

The widespread use of standard computer packages has encouraged the search for simple modifications to standard test procedures to take into account complex sample design. The idea can be regarded as a natural extension of the use of design effects as multiplicative factors for variances based on a simple random sample of the same size, in order to correct for the complex design used.

The correction may indeed be based on design effects of various estimators or on average design effects (see, e.g., Cowan and Binder (1978), Fay (1979), Fellegi (1980), Rao and Scott (1981) and Scott and Holt (1981).

Another alternative is to investigate the behaviours of standard test statistics under some superpopulation model and to modify the standard statistic accordingly (Cohen (1976) and Campbell (1977)).

## 3. SPECIFIC METHODS

### 3.1 Linear Models and Regression

The prior determination of the model and of the parameters of interest is extremely important for the case of regression analysis and of linear models. For instance, when different regression relationships must be assumed for different strata or for different PSU's in a two-stage design, the parameter of interest could be a simple average of the regression coefficients (Konijn (1962)); a weighted average of the coefficients (Pfeffermann and Nathan (1981)); or their expected value (under some prior distribution) (Porter (1973)).

The model and the parameters of interest should, in general, be determined on the basis of the assumed overall population structure and should not reflect to the structure of the sample design. However in many cases the sample design will reflect population structure so that sample design variables may be part of the model. For example consider the model:

$$E(Y|X_1,X_2) = X_1\ \beta_{1.2} + X_2\ \beta_{2.1} \tag{3.1.1}$$

where $X_1$ includes only variables which do not relate to the sample design and $X_2$ includes all the variables which enter into the complex sample design, i.e. the sample distribution depends only on $X_2$:

$$P(s|X_1,X_2) = P(s|X_2). \tag{3.1.2}$$

The estimation of $\beta_{1.2}$ and of $\beta_{2.1}$ in (3.1.1) and inference about them can proceed in the classical way, as if sampling were simple random, if indeed (3.1.1) holds.

However if the design variables, $X_2$, are not included in the regression equation of interest:

$$E(Y|X_1) = X_1\beta_1 \tag{3.1.3}$$

and the design variable $X_2$ is correlated with Y (conditional on $X_1$) then the standard OLS estimator of $\beta_1$ is not consistent (see Nathan and Holt (1980) and Holt and Smith (1979), who propose modified weighted and unweighted estimates of $\beta_1$, which are consistent). Holt, Smith and Winter (1980) give an example of the application of these estimators.

If the linear model:

$$E(Y_i|x_i) = x_i'\ \beta \tag{3.1.4}$$

$$\operatorname{cov}(Y_i,Y_j|x_i,x_j) = \begin{cases}\sigma^2 & i=j\\ 0 & i\neq j\end{cases} \tag{3.1.5}$$

indeed holds for all population units (i, j=1, ..., N) of a finite population and the p×1 column vector $x_i$ includes all the sample design variables, then the OLS unweighted estimator:

$$\hat{\beta} = (X_n'\ X_n)^{-1}\ X_n'\ Y_n \tag{3.1.6}$$

based on the sampled values $\underset{p\times n}{X_n'} = (x_1, \ldots, x_n)$ and $Y_n' = (Y_1, \ldots, Y_n)$ is the "best" linear model-unbiased estimator of $\beta$ irrespective of the sample design. "Best" here is in the sense of minimal model-variance. However $\hat{\beta}$ is, in general, not a design-unbiased, nor even a design-consistent, estimator of the population parameter:

$$B = (X_N' X_N)^{-1} X_N' Y_N , \qquad (3.1.7)$$

where $\underset{p\times N}{X_N'} = (x_1, \ldots, x_N)$ and $Y_N' = (Y_1, \ldots, Y_N)$.

The design-consistent estimator of B is the weighted estimator:

$$\hat{\beta}_W = (X_n' W_n X_n)^{-1} X_n' W_n Y_n , \qquad (3.1.8)$$

where the weight matrix, $W_n = \text{diag}\ (\Pi_1^{-1}, \ldots, \Pi_n^{-1})$, is the n×n diagonal matrix of the reciprocals of the sample inclusion probabilities $\Pi_i = \Pr(i\varepsilon s)$.

The consistency of $\hat{\beta}_W$, as an estimator of B, obviously does not depend on the model (3.1.4) holding, but the relevance of estimating B when the model does not hold can be challenged. It can be shown that under certain conditions for a non-linear model, which assumes that the conditional expectation of Y (given X) is a differentiable function of X, the model-expectation of B can be expressed approximately as a weighted average of the slopes of this function at the points $X_i$ (the weights depending only on $X_i - \bar{X}$). However this interpretation is of limited practical value.

In any case $\hat{\beta}_W$ is a model-unbiased estimator of $\beta$, whenever (3.1.4) does hold. It will not, in general, be an optimal estimator of $\beta$ under (3.1.5) for unequal probability sampling, but will be so if the conditional model variance of $Y_i$ is proportional to $\Pi_i$, i.e.

$$V(Y_i|x_i) = k\,\Pi_i \; . \qquad (3.1.9)$$

Since the weighted estimator, $\hat{\beta}_W$, is more robust then the unweighted estimator, $\hat{\beta}$, in the sense that it is both a model-unbiased estimator of $\beta$, if the model holds and a design-consistent estimator of B, if not, the use of the weighted estimator $\hat{\beta}_W$ is recommended, for estimation of B, whenever there is no assurance that the model (3.1.4)-(3.1.5) holds. The question which must then be answered by the subject-matter specialist is whether B is a relevant parameter to estimate.

It should be noted that for self-weighting designs $\hat{\beta}$ and $\hat{\beta}_W$ coincide. The estimator, $\hat{\beta}_W$ (3.1.8),can be obtained directly from standard computer programmes which provide for weighted regression (e.g.BMDP) by using the weights $1/\Pi_i$; or from other programmes (e.g. SPSS) by carrying out unweighted regression on the transformed variables $Y_i/\sqrt{\Pi_i}$ and $x_i/\sqrt{\Pi_i}$, but not on the weighted variables $Y_i/\Pi_i$, $x_i/\Pi_i$. However, it should be noted that under either alternative the reported variances and covariances of the estimators are incorrect and that the standard significance tests (e.g. F tests) are invalid, and can result in grossly misleading conclusions.

Assuming the model (3.1.4) - (3.1.5), the model variance of $\hat{\beta}$ is:

$$V(\hat{\beta}|X_n) = \sigma^2(X_n' X_n)^{-1} \; , \qquad (3.1.10)$$

which is the result given by standard unweighted regression programmes. However, the model variance of $\hat{\beta}_W$ is:

$$V(\hat{\beta}_W|X_n) = \sigma^2(X_n' W_n X_n)^{-1} X_n' W_n' W_n X_n (X_n' W_n X_n)^{-1} \; . \qquad (3.1.11)$$

The weighted regression programme, with weights $1/\Pi_i$, will give a value of $(X_n' W_n X_n)^{-1}$ for the model variance of $\hat{\beta}_W$, which equals (3.1.11) only if $W_n = I_n$. Thus none of the standard outputs for standard errors or for tests of hypotheses are correct.

However the estimator of the multiple correlation coefficient obtained from weighted regression:

$$\hat{R}^2 = \frac{(Y_n - X_n \hat{\beta}_W)' W_n (Y_n - X_n \hat{\beta}_W)}{(Y_n - \bar{y}_n \underline{1}_n)' W_n (Y_n - \bar{y}_n \underline{1}_n)} , \qquad (3.1.12)$$

where $\bar{y}_n = (\Sigma_s Y_i/\Pi_i) / (\Sigma_s 1/\Pi_i)$, is a design-consistent estimator of the population multiple correlation coefficient:

$$R^2 = \frac{(Y_N - X_N B)' (Y_N - X_N B)}{(Y_N - \bar{Y}_N \underline{1}_N)' (Y_N - \bar{Y}_N \underline{1}_N)} \qquad (3.1.13)$$

where $\bar{Y}_N = (1/N) \underline{1}_N' Y_N$.

The design-variance of $\hat{\beta}_W$, which must be considered the relevant measure of accuracy for $\hat{\beta}_W$ as an estimator of B, cannot in general, be obtained from only the first order inclusion probabilities, $\Pi_i$. For most sample designs used in practice, the design-variance of $\hat{\beta}_W$ will have to be estimated by one of the variance estimating techniques mentioned above i.e.linearization, Balanced Repeated Replication or Jackknife (see, e.g., Jonrup and Remmermalm (1976) and Holt and Scott (1981)).

## 3.2 Categorical Data Analysis

The simplest analysis of categorical data relates to a single classification of the population into k classes with probabilities (relative frequencies) $\underline{p}' = (p_1, \ldots, p_{k-1})$. In order to test the null hypothesis of goodness of fit to a known distribution $\underline{p}_o' = (p_{o1}, \ldots, p_{ok-1})$:

$$H_o: \quad \underline{p} = \underline{p}_o , \qquad (3.2.1)$$

the approaches outlined in section two can be used.

We assume that a consistent survey estimator $\hat{\underline{p}}' = (\hat{p}_1, \ldots, \hat{p}_{k-1})$ of $\underline{p}'$ is available. If it is asymptotically normal:

$$\sqrt{n}\ (\hat{\underline{p}} - \underline{p}) \rightarrow N(\underline{0},V) \tag{3.2.2}$$

and a consistent estimator, $\hat{V}$, of V is available, then the generalized Wald statistic:

$$X_W^2 = n(\hat{\underline{p}} - \underline{p}_o)'\ \hat{V}^{-1}\ (\hat{\underline{p}} - \underline{p}_o)\ , \tag{3.2.3}$$

which is distributed asymptotically as $\chi^2_{k-1}$ under $H_o$, can be used to test $H_o$.

For many simple designs consistent estimators of V are directly available and for more complex designs they can be obtained by standard methods. However if tests of hypotheses of goodness of fit have to be carried out for a variety of variables and classifications, the use of the standard $X^2$ statistic:

$$X^2 = n \sum_{i=1}^{k} (\hat{p}_i - p_{oi})^2/p_{oi} = n(\hat{\underline{p}} - \underline{p}_o)'\ P_o^{-1}(\hat{\underline{p}} - \underline{p}_o)\ , \tag{3.2.4}$$

where $P_o = \text{diag}\ (\underline{p}_o) - \underline{p}_o\,\underline{p}_o'$, with appropriate modification may be prefered. Rao and Scott (1981) show that the asymptotic distribution of $X^2$ under $H_o$ is that of a weighted sum of k-1 independent $\chi^2$ variables with one degree of freedom each.

$$X^2 \rightarrow \sum_{i=1}^{k-1} \lambda_i\ Z_i^2;\quad Z_i \sim N(0,1)\quad \text{independent} \tag{3.2.5}$$

where $\lambda_1, \ldots, \lambda_{k-1}$ are the eigenvalues of

$$D = P_o^{-1}\ V\ (\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{k-1} > 0). \tag{3.2.6}$$

A conservative test of (3.2.1) can then be obtained by using the statistic $X^2/\lambda_1$ in conjunction with a $\chi^2_{k-1}$ distribution. $\lambda_1$ can be components of $\hat{\underline{p}}$. For example, for proportional stratified sampling $\lambda_1 \leq 1$, so that $X^2$ itself can be used as a conservative test statistic.

In other cases the use of $X^2/\bar{\lambda}$ with:

$$\bar{\lambda} = \frac{1}{k-1}\sum_{i=1}^{k-1}\lambda_i = \frac{1}{k-1}\sum_{i=1}^{k} d_i(1-p_i) ,$$

where $d_i = V[\hat{p}_i]/[p_i(1-p_i)]$ is the design effect for $\hat{p}_i$, has been shown to be a good approximative test by Hidiroglou and Rao (1981) for the Canada Health Surveys and by Holt, Scott and Ewings (1980) for large scale U.K. surveys. An alternative approximation - $X^2/\bar{d}$, where $\bar{d} = k^{-1}\sum_{i=1}^{k} d_i$ - has been proposed by Fellegi (1980).

Direct modelling for $\underline{p}$ has been proposed by Altham (1976) and by Cohen (1976), but their models have the serious limitation that they imply $\lambda_1 = \lambda_2 = \ldots = \lambda_{k-1} = \bar{\lambda}$, which is equivalent to a constant design effect over categories. This is not a realistic assumption, in general, and results in $X^2/\bar{\lambda}$ having exactly an asymptotic $\chi^2_{k-1}$ distribution.

For testing independence in a two-way contingency table, the hypotheses can be formulated:

$$H_o:\ h_{ij}(\underline{p}) = p_{ij} - p_{i+}\,p_{+j} = 0$$

$$(i=1,\ \ldots,\ r-1;\ j-1,\ \ldots,\ c-1), \qquad (3.2.7)$$

where $p_{ij}$ is the population probability of cell $(i,j)$ $p_{i+}$, $p_{+j}$ are the marginal probabilities and $\underline{p}' = (p_{11}, \ldots, p_{rc-1})$. The generalized Wald statistic for testing $H_o$ is:

$$X^2_{W1} = n[\underline{h}(\hat{\underline{p}})]'\ \hat{V}_h^{-1}\ \underline{h}(\hat{\underline{p}}) , \qquad (3.2.8)$$

where $[\underline{h}(\hat{\underline{p}})]' = [h_{11}(\hat{\underline{p}}), \ldots, h_{r-1\ c-1}\ (\hat{p})]$ and $\hat{V}_h/n$ is a consistent estimator of the covariance matrix of $\underline{h}(\hat{\underline{p}})$. Versions of (3.2.8) for specific designs with various methods for estimating $\hat{V}_{h/n}$ have been used by Garza-Hernandez and McCarthy (1962), Nathan (1969, 1975) Shuster and Downing (1976) and Fellegi (1980).

A modified statistic similar to $X^2/\bar{\chi}$ has been proposed by Rao and Scott (1981):

$$X^2_{CI} = (n/\hat{\delta}) \sum_{i=1}^{r} \sum_{j=1}^{c} (\hat{p}_{ij} - \hat{p}_{i+} \hat{p}_{+j})^2/(\hat{p}_{i+} \hat{p}_{+j}), \qquad (3.2.9)$$

where $\hat{\delta} = \frac{1}{(r-1)(c-1)} \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{v}_{ij}(\underline{h})/(\hat{p}_{i+} \hat{p}_{+j})$ and

$\hat{v}_{ij}(\underline{h})/n$ is an estimator of the variance of $h_{ij}(\hat{\underline{p}})$. $\hat{\delta}$ can be written in terms of the estimated deffs of $h_{ij}(\hat{\underline{p}})$:

$$\hat{\delta} = \frac{1}{(r-1)(c-1)} \sum_{i=1}^{r} \sum_{j=1}^{c} (1 - \hat{p}_{i+})(1 - \hat{p}_{+j})\, \hat{\delta}_{ij}, \qquad (3.2.10)$$

where $\hat{\delta}_{ij}$ is an estimator of the deff, $\delta_{ij}$, of $h_{ij}(\hat{\underline{p}})$ :

$$\delta_{ij} = nV[h_{ij}(\hat{\underline{p}})] \;/\; [p_{i+}\, p_{+j}\, (1 - p_{i+})(1 - p_{+j}) \;. \qquad (3.2.11)$$

Estimates of the design effects may be easier to obtain than estimates of variances.

Empirical investigations by Holt, Scott and Ewings (1980) and by Hidiroglou and Rao (1981) indicate that the distribution of $X^2_{CI}$ is close to $\chi^2_{(r-1)(c-1)}$.

## 3.3 Other Types of Analysis

While linear models, tests of goodness of fit and tests of independence cover many important analysis applications, other types of analysis, such as principal component and factor analysis,discriminant analysis, path analysis, logistic regression, log-linear models non-parametric methods, etc. cannot be directly dealt with in the same way. While the general techniques outlined in section two could be used, their application presents difficulties and only few cases of their application have been reported.

Since correlation coefficients are a basic element in most multivariate analysis, some empirical studies of the effect of sample design on their estimation have been carried out by Kish and Frankel (1974), Bebbington and Smith (1977) and Holt, Richardson and Mitchell (1980). No general conclusions can be formulated, but design effects are definitely not negligible. Bebbington and Smith (1977) have also studied the sampling variability of principal components estimators.

In other areas design effects for logits have been studied by Lepkowski and Landis (1980) and confidence intervals for quantiles by Woodruff (1952) and by Sedransk and Meyer (1978).

## ACKNOWLEDGEMENTS

This paper has benefited from comments by and dicussion with D. Binder, N. Chinnappa, S.E. Fienberg, M. Hidiroglou, G.J.C. Hole, J.N.K. Rao and A. Scott.

## REFERENCES

[1] Altham, P.M.E. (1976), "Discrete Variable Analysis for Individuals Grouped Into Families", Biometrika, 63, 263-269.

[2] Brewer, K.R. and Mellor, R.W. (1973), "The Effect of Sample Structure on Analytical Surveys", Aust. J. Statist., 15, 145-152.

[3] Bebbington, A.C. and Smith, T.M.F. (1977), "The Effect of Survey Design on Multivariate Analysis", The Analysis of Survey Data (C.A. O'MUIRCHEARTAIGH and C. PAYNE, EDITORS). Vol. 2, Model Fitting, New York: Wiley, 175-192.

[4] Campbell, C. (1977), "Properties of Ordinary and Weighted Least Squares Estimators for Two Stage Samples", Proc. Soc. Statist. Sect., Ameri. Statist. Assoc., 800-805.

[5] Cohen, J.E. (1976), "The Distribution of the Chi-Squared Statistic Under Clustered Sampling", J. Amer. Statist. Assoc. 71, 665-670.

[6] Cowan, J. and Binder, D.A. (1978). "The Effect of a Two-Stage Sample Design on Tests of Independence", Survey Methodology, Vol. 4, No. 1, 16-29.

[7] Fay, R.E. (1979), "On Adjusting the Pearson Chi-Square Statistic for Clustered Sampling", Proc. Soc. Statist. Sect., Amer. Statist. Assoc. 402-405.

[8] Fellegi, I.P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples", J. Amer. Statist. Assoc. 75, 261-268.

[9] Fienberg, S.E. (1980), "The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey", The Statistician, 29, 313-350.

[10] Fuller, W.A. (1975), "Regression Analysis for Sample Survey", Sankhya C, 37, 117-132.

[11] Fuller, W.A. and Battese, G.E. (1973), "Transformations for Estimation of Linear Models with Nested-Error Structure", J. Amer. Statist. Assoc. 68, 626-632.

[12] Fuller, W.A. and Rao, J.N.K. (1978), "Estimation for a Linear Regression Model with Unknown Diagonal Covariance Matrix", Ann. Statist. 6. 1149-1158.

[13] Freeman, D.H. Jr., Freeman, J., Brock, D.B. and Koch, G.G., "Strategies in the Multivariate Analysis of Data from Complex Surveys II: An Application to the United States National Health Interview Survey", Inter. Statist. Rev. 44, 317-330.

[14] Garza-Hernandez, T. and McCarthy, P.J. (1962), "A Test of Homogeneity for a Stratified Sample", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 200-202.

[15] Grizzle, J.E., Starmer, C.F. and Kock, G.G. (1969), "Analysis of Categorical Data by Linear Models", Biometrics, 25, 489-504.

[16] Hartley, H.O. and Sielken, R.L. (1975), "A Superpopulation Viewpoint for Finite Population Sampling", Biometrics, 31, 411-422.

[17] Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1980). Super Carp: Sixth Edition, Statistical Laboratory Survey Section, Iowa State University, Ames, Iowa.

[18] Hidiroglou, M.A. and Rao, J.N.K. (1981), "Chisquare Tests for the Analysis of Categorical Data from the Canada Health Survey", Invited Paper for 43rd Session of I.S.I., Buenos-Aires.

[19] Holt, D., Richardson, S.C. and Mitchell, P.W. (1980), "The Analysis of Correlations in Complex Survey Data", (unpublished).

[20] Holt, D. and Scott, A.J. (1981), "Regression Analysis using Survey Data", The Statistician, 30. (to appear).

[21] Holt, D., Scott, A.J., and Ewings, P.O. (1980), "Chi-Squared Tests with Survey Data", J. Roy. Statist. Soc. A., 143, 302-330.

[22] Holt, D., Smith, T.M.F. (1979), "Regression Analysis of Data from Complex Surveys", Roy. Statist. Soc. Conf., Oxford.

[23] Holt, D., Smith, T.M.F. and Winter, P.O. (1980), "Regression Analysis of Data from Complex Surveys", Jour. Roy. Statist. Soc. A, 143, 474-483.

[24] Imvrey, P., Sobel, E. and Francis, M. (1980), "Modeling Contingency Tables from Complex Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 213-217.

[25] Jonrup, H. and Rennermalm, B. (1976), "Regression Analysis in Samples from Finite Population", Scand. Jour. Statist., 3, 33-37.

[26] Kaplan, B., Francis, I., and Sedransk, J. (1979), "A Comparison of Methods and Programs for Computing Variances of Estimators from Complex Sample Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 97-100.

[27] Kish, Leslie and Frankel, M.R. (1970), "Balanced Repeated Replication for Standard Errors", J. Amer. Statist. Assoc., 65, 1071-1094.

[28] Kish, L. and Frankel, M.R. (1974), "Inference from Complex Samples (with discussion)", J. Roy. Statist. Soc. B, 36, 1-37.

[29] Koch, G.G., Freeman, D.H., Jr., and Freeman, J.L. (1975), "Strategies in the Multivariate Analysis of Data from Complex Surveys", Inter. Statist. Rev. 43, 59-78.

[30] Koch, G.G., Stokes, M.E. and Brock, D. (1980), "Applications of Weighted Least Squares Methods for Fitting Variational Models to Health Survey Data", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 218-223.

[31] Konijn, H.S. (1962), "Regression Analysis for Sample Surveys", J. Amer. Statist. Assoc. 57, 590-606.

[32] Krewski, D., and Rao, J.N.K. (1981), "Inference from Startified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods", Ann. Statist., 9 (5) 1010-1019.

[33] Lepkowski, J.N. and Landis, J.R. (1980), "Design Effects for Linear Contrasts of Proportions and Logits", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 224-229.

[34] McMarthy, P.J. (1969). "PSEUDO-REPLICATION: Half-Samples", Inter Statist. Rev. 37, 239-264.

[35] Miller, R.G. (1974), "The JACKKNIFE- A Review", Biometrika 61, 1-15.

[36] Nathan, G. (1969), "Tests of Independence in Contingency Tables from Stratified Samples", New developments in Survey Sampling (N.L. Johnson and H. Smith, eds.). New York: Wiley, 578-600.

[37] Nathan, G. (1972), "On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples", J. Amer. Statist. Assoc., 67, 917-920.

[38] Nathan, G. (1973), "Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples", National Center for Health Statistics, Vital and Health Statistics Series 2, No. 53, Washington, D.C.

[39] Nathan, G. (1975), "Tests of Independence in Contingency Tables from Stratified Proportional Samples", Sankhya C, 37, 77-87. [corrigendum: Sankhya C, 40, (1978), 190].

[40] Nathan, G. and Holt, D. (1980), "The Effect of Survey Design on Regression Analysis", J. Roy. Statist. Soc. B, 42, 377-386.

[41] Pfeffermann, D., and Nathan, G. (1981), "Regression Analysis of Data from Complex Samples", J. Amer. Statist. Assoc., 76, 681-689.

[42] Porter, R.M. (1973), "On the Use of Survey Sample Weights in the Linear Model", Annals of Economic and Social Measurement, 2, 141-158.

[43] Rao, J.N.K. (1975), "Analytic Studies of Sample Survey Data", Survey Methodology, Vol. 1, Supplementary Issue.

[44] Rao, J.N.K. and Scott, A.J. (1981), "The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables", J. Amer. Statist. Assoc. 76, 221-230.

[45] Richards, V. and Freeman, D.H. Jr. (1980), "A Comparison of Replicated and Pseudo-Replicated Covariance Matrix Estimators for the Analysis of Contingency Tables", Proc. Sec. Survey Meth., Amer. Statist. Assoc., 209-211.

[46] Särndal, C.E. (1978), "Design-Based and Model-Based Inference in Survey Sampling", Scand. J. Statist., 5, 27-52.

[47] Sedransk, S. and Meyer, J. (1978), "Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling", J. Roy. Statist. Soc. B., 40, 239-252.

[48] Scott, A. and Holt, D. (1981), "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods", (unpublished)

[49] Shah, B.V. (1978), "SUDAAN: Survey Data Analysis Software", Proc. Statist. Comp. Sect., Amer. Statist. Assoc., 146-151.

[50] Shah, B.V., Holt, M.M. and Folsom, R.E. (1977), "Inference about Regression Model from Sample Survey Data", Bull. Inter. Statist. Inst. 47, Bk. 3, 43-57.

[51] Shuster, J.J. and Downing, D.J. (1976), "Two-Way Contingency Tables for Complex Sampling Schemes", Biometrika 63, 271-278.

[52] Smith, T.M.F. (1976), "The Foundations of Survey Sampling: A Review (with discussion)", J. Roy. Statist. Soc. A., 139, 183-195.

[53] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 66, 411-414.

[54] Thomsen, I. (1978), "Design and Estimation Problems when Estimating a Regression Coefficient from Survey Data", Metrika 25, 27-35.

[55] Tomberlin, T.J. (1979), "The Analysis of Contingency Tables of Data from Complex Samples", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 152-157.

[56] Woodruff, Ralph S. (1952), "Confidence Intervals for Medians and Other Position Measures", J. Amer. Statist. Assoc. 47, 635-646.

# THE NONRESPONSE PROBLEM

J.G. BETHLEHEM AND H.M.P. KERSTEN[1]

This paper presents an outline of the nonresponse research which is carried out at the Netherlands Central Bureau of Statistics. The phenomenon of nonresponse is put into a general frame-work. The extent of nonresponse is indicated with figures from a number of CBS-surveys. The use of auxiliary variables is discussed as a means for obtaining information about nonrespondents. These variables can be used either to characterize nonrespondents or as stratification variables in adjustment procedures.

Adjustment for nonresponse bias by means of subgroup weighting is considered in more detail. Finally, the last section lists a number of other methods which also aim at reduction of the bias.

## 1. INTRODUCTION

Nonresponse is becoming a growing concern in survey research. The phenomenon of nonresponse, when people are not able or willing to answer questions asked by the interviewer, can appear in sample surveys as well as in censuses. It affects the quality of the survey in two ways: first of all, due to reduction of the available amount of data, estimates of population parameters will be less precise. Secondly, if a relationshiɔ exists between the variable under investigation and response behaviour, statements made on the basis of the response are not valid for the total population. For example if the housing demand of respondents is greater than the housing demand of nonrespondents, estimates of the housing demand in the total population will be significantly too high.

[1] J.G. Bethlehem and H.M.P. Kersten, Netherlands Central Bureau of Statistics.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Netherlands Central Bureau of Statistics.

It is obvious that the extent of the nonresponse must be kept as small as possible. If, in spite of these efforts, there still remains a considerable amount of nonresponse, measures have to be taken in order to prevent formulation of wrong statements about the population. Combination of adjustment procedures and usual estimation techniques is necessary to yield valid population estimates.

Two departments of the CBS (Netherlands Central Bureau of Statistics) are involved in nonresponse research. The Department for Social Surveys is responsible for the field work of the surveys. It is concerned with minimizing nonresponse during the process of collecting data. Research is carried out on the optimal number of recalls and the time of the interview. (See Widdershoven & Van den Berg (1980).) Experiments are set up to find the optimal way to approach persons and households with introductory letters. Attempts are made to measure the impact of interview fatigue and interview pressure. Ultimately, notwithstanding these efforts, there still remains an amount of nonresponse. The Department for Statistical Methods investigates the effect of nonresponse on the accuracy of the results of the survey. Methods are developed there to adjust population estimates for the bias due to nonresponse. The remainder of this paper is mainly concerned with the work of the latter department.

The next sections present an outline of the nonresponse analysis at the CBS. Section 2 introduces definitions and the accompanying problems. Nonresponse figures of a number of CBS-surveys are summarized. In section 3 graphical methods are discussed to select auxiliary variables. They provide insight into nonresponse and can be used in adjustment procedures. Section 4 presents adjustment methods which make use of subgroup weighting and section 5 lists a number of other methods.

## 2. THE PHENOMENON OF NONRESPONSE

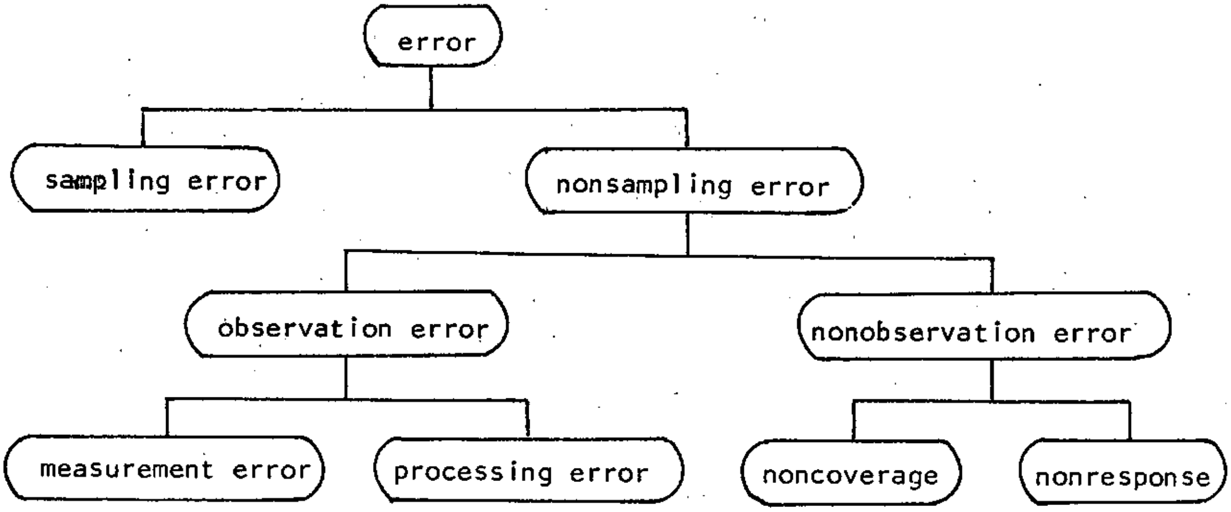
In this section the problem of nonresponse is placed in a general framework, in which also a number of other sampling problems play a role.

Nonresponse figures for a number of CBS surveys are given. Situations are described in which a relationship exists between the variable under investigation and the response behaviour. In the last part of the section two models for the general of nonresponse are considered.

## 2.1 Terminology

The objective of every survey is the determination of certain population characteristics. Due to all kinds of errors, the true value will generally never be obtained. A typology of sources of error is presented in fig. 1. The scheme is due to Kish (1967).

FIG. 1. TYPOLOGY OF ERRORS IN SURVEYS



The two sources of error in surveys are sampling errors and nonsampling errors. Sampling errors consist of that part of the error which is due to the fact that only a sample of values is observed rather than the total population. The sampling error has an expected frequency distribution generated by the totality of sampling errors in all possible samples of the same size. This distribution is used to estimate the population characteristic.

Nonsampling errors are those errors in sample estimates which can not be attributed to sampling fluctuations. Nonsampling errors are often a more serious problem than sampling errors. Nonsampling errors can be divided in observation errors and nonobservation errors.

Observation errors are caused by obtaining and recording observations incorrectly. They may be further subdivided into measurement errors and processing errors.

Measurement errors are caused either by the interviewer or by the respondent. The interviewer himself can be a source of error. He can influence the response by his mere presence, by his (or her) sex, skin colour, age, or dress. Also the way in which he asks questions and clarifies statements affects results. The answer of a person may depend on the type of question (whether a question measures a fact such as year of birth, or an opinion). Errors can also be introduced by factors such as whether the person understands the question, whether he knows the answer or not, whether he wishes to conceal the answer, or whether he wishes to present a certain image. Moreover, memory is not always free of errors, and data may be incorrectly recorded.

Processing errors arise during the processing of the data at the office. They occur during the stage of coding, tabulating and computing.

Nonobservation errors are due to the failure to obtain observations on certain parts of the population. They may be subdivided in noncoverage and nonresponse.

Let the target population be the population the survey is intended to cover. Practical difficulties in handling parts of the population may result in their elimination from the scope of the survey. It is also possible that the actually sampled population contains elements which do not belong to the scope of the survey.

Noncoverage refers to all errors which result from differences between target population and sampled population. Elements which belong to the target population as well as to the sampled population are correct elements. The situation in which elements in the target population do not appear in the sampled population is called undercoverage. These elements have zero probability of selection in the sample. The situation in which elements in the sampled population do not appear in the target population is called overcoverage. Elements, classified as overcoverage, are called duds. They have to be excluded from the sample before analysis takes place. If there is unexpected overcoverage the ultimate sample size may be less than the planned sample size.

Nonresponse refers to failure to obtain observations on some elements selected and designated for the sample. A good classification of nonresponse errors depends on the survey situation. The classification given below focuses on problems in face-to-face interviews. A similar treatment may be applicable in other survey situations. The following categories of nonresponse can be distinguished:

(1) Not at home. To reduce the extent of this category recalls can be made. Research should be carried out on the optimal number of recalls. The term temporarily unavailable would be a useful generalization for this category, denoting a delay rather than a denial of the interview. The respondent may be too busy, tired, or ill at the time, but will be cooperative on another call.

(2) Refusal. Some of the factors causing refusal are temporary and changeable. A person may refuse because he is ill-disposed or approached at the wrong hour. Another try, or another approach may find him cooperative. Since quite a number of refusals can, however be considered permanent, a better term for this category is unobtainable,, denoting a denial rather than a delay of observation. Repeated attempts will not bring success. From this view, respondents known to be away during the entire survey period belong in this category, rather than among the not-at-homes.

(3) Incapacity or inability. This type of nonresponse may refer to mental or physical illness which prevents response during the entire survey period. A language barrier belongs also to this category. If generalized this category could fit in the previously defined unobtainables. It can, however, be useful in some situations to distinguish between the unwilling and the willing, but incapable, respondent.

(4) Not found. This category can e.g. be large for movers. Such respondents are either not identified or followed because this would be too expensive. Cases of not attempted interviews belong to the same general category. They could be caused by inaccessibility (lighthouse keeper, shepherd), or dangerous surroundings (watchdog, slum).

(5) Lost information. Information may get lost after a field attempt. Some questionnaires may be unusable because of poor quality or cheating. Other may remain unfilled because they were lost or forgotten.

The typology as described above is applicable in most survey situations, but care must be taken in case of complex sampling designs. When e.g. sampling takes place in more stages the typology can be used in each separate stage. The same source of error can be classified differently in different stage. This is illustrated in an example. In a household survey first a sample of households is selected. The interviewer enumerates all persons in a particular selected household and after that selects a sample from this list. In such an enumeration the student living in an attic is often concealed. In the first stage of the sampling procedure this situation would be classified as measurement error, and in the second stage as undercoverage.

For some sources of error classification may depend on other factors and appropriate rules to cover them must be adopted. For example, if a person to be interviewed died before the interview could take place, classification depends on the time of death. If death occurred before the day the sample was selected this could be classified as overcoverage, but if death occurred between the day the sample was selected and the day of the interview, the correct classification may be nonresponse.

Before selecting the sample, the population must be divided into sampling units. To every element in the population there must correspond one and only one sampling unit. The construction of the physical list of sampling units, called the sampling frame, is often a major practical problem. The nature of the available sampling frames is an important consideration in sample design. Relevant factors include the type of sampling unit, extent of coverage, accuracy and completeness of the list, and the amount and quality of auxiliary information in the list.

For sampling frames in which the sampling unit is a person the CBS has to restrict itself to administrative records of local authorities (municipalities). For household surveys the CBS manages its own frame, but at the moment the use of the list of delivery points of the Post Office is considred as a sampling frame.

## 2.2 The Extent of Nonresponse

It is rather difficult to compare nonresponse figures of different surveys. The percentage of nonresponse depends on a number of circumstances: aim of the survey, type of sampling unit, the sampling design, efficiency of the field work, performance of the interviewers, nonresponse reducing measures, perdiod in which the survey is held, the target population, the length of the questionnaire, wording of questions, etc. Even the definition of nonresponse may differ. It is necessary to create a frame-work which enables proper comparison of surveys. By controlling the factors which influence nonresponse figures, judgement can be passed on the quality of the different surveys. Such a frame work also offers opportunities for comparing surveys from different countries.

Table 1 presents nonresponse figures of a number of CBS-surveys. A clear trend of increasing nonresponse percentages can be seen in this table.

Table 1: Nonresponse percentages of some CBS-surveys

| year | LFS | | SSC | | SLC | | NTS | | HS | |
|---|---|---|---|---|---|---|---|---|---|---|
| | tn | rn | tn | rn | tn | rn | tn | rn | tn | rn |
| 1973 | 13.2 | | | | | | | | | |
| 1974 | | | | | 28.2 | 15.6 | | | | |
| 1975 | 15.8 | 9.0 | 30.1 | 18.3 | | | | | 14.5 | |
| 1976 | | | 28.1 | 18.6 | 23.0[1] | 15.6 | | | 12.9 | |
| 1977 | 13.1 | 6.6 | 30.9 | 20.5 | 29.7 | 16.9 | | | 17.6 | 9.3 |
| 1978 | | | 36.1 | 23.9 | | | 33.0 | 26.2 | 21.9 | 12.5 |
| 1979 | 19.7 | | 36.6 | 24.4 | 33.7[2] | | 30.6 | 23.9 | 25.5 | |
| 1980 | | | 36.8 | 24.7 | 35.6 | 19.7 | 32.1 | 24.5 | | |

1) = elderly people only

2) = young people only

tn = percentage of total nonresponse

rn = percentage of refusals

LFS = Labour Force Survey

SSC = Survey of Consumer Sentiments

SLC = Survey of Living Conditions

NTS = National Travel Survey

HS = Holiday Survey

As mentioned before a relationship between the variable under investigation and the response behaviour reduces the value of the conclusions of the survey. The existence of such relationships is not rare, as will be illustrated in the following examples. If the aim of the survey is to measure in which way people spend their spare time, then the reason of nonresponse "not at home" is rather annoying since these people are probably spending their (spare) time somewhere else. The same applies for the survey on the number of hours people watch television: the not-at-homes (in the evening) are probably not watching television. One of the aims of the Housing

Demand Survey is to measure the frequency with which people move to other houses. As there is a considerable amount of nonresponse due to moving (the sampling unit is a person), the estimate for the total population will be biased. A number of surveys show that unmarried people have a smaller response rate. If there is a relationship between marital status and the variable under investigation then estimates will be wrong in this case too.

## 2.3 Response Models

The first requirement in the development of theories for the treatment of nonresponse is the formulation of a mathematical model, which describes the way in which nonresponse is generated. Two models appear frequently in the literature. They are denoted here by "random response model" and "fixed response model".

According to the random response model every element in the population has a certain (unknown) probability of response. These response probabilities are not necessarily the same for every element. When the interviewer contacts the person to be questioned the probability mechanism is activated and determines whether or not the person responds.

The fixed response model assumes the existence of two strata in the population: a stratum of potential respondents and a stratum of potential nonrespondents. Size and content of each stratum is not known beforehand. They are determined by the specification of the survey (aim, type of questions, interviewing techniques, interviewers, period of field work, etc.). Disregarding the two strata a sample is selected from the population. Consequently the number of respondents is a random variable in both the random response model and the fixed response model.

If instead of sampling complete enumeration would take place then in the case of random response model the determination of respondents would still be a random process whereas in the case of the fixed response model this would be fixed. There is, however, a certain resemblance between the two models. Assuming the existence of two stochastic meachanisms, the sampling mechanism and the response mechanism, both models differ only in the order in which the mechanisms are applied: In the fixed response model first the response mechanism is activated for each element in the population. This determines the two strata. Then the sample is selected. In the random response model first the sample is selected. Then the response mechanism is activated for each selected element.

The random response model offers the opportunity to estimate response probabilities. These estimated response probabilities can be used in adjustment procedures, or they can be connected to personal characteristics. The fixed response models generally results in easier formulae. The theory, developed within this model, is conditional on the realized response and non-response strata. Consequently the accuracy of the estimates can be computed, but the accuracy of the estimation method can not be determined. Due to this last argument research is focussed on the random response model.

## 3. SELECTION OF AUXILIARY VARIABLES

### 3.1 Auxiliary Variables

It is important to discover a possibly existing relationship between the variable under investigation and the response behaviour. It is, however, not possible to determine such a relationship using the sample data, since the values of the variable under investigation are not known for the nonrespondents. To be able to say something about nonrespondents there must be information available about them. One source of information about the nonresponse is formed by auxiliary variables. Auxiliary variables are defined as variables which can be measured for both respondents and nonrespondents. Two types of auxiliary information can be distinguished:

(1) Information which can be collected by the interviewer without a face-to-face interview. Among the information, obtained in this way, are type of town, type of housing, (approximate) year of construction of the housing and social status of the neighbourhood.

(2) Information which can be obtained from administrative records. Typical examples are age, sex and marital status.

Analysis of the relationship between auxiliary variables and the response behaviour provides insight in the group of people which do not respond. It may give additional information about the relationship between the variable under investigation and the response behaviour. Auxiliary variables showing a clear relationship with the response behaviour play an important rôle in adjustment procedures, to be discussed later.

It is assumed that auxiliary variables are nominal variables, i.e. different values have no other meaning than to distinguish between different groups. Arithmetic operations on these values, which in fact are only labels, are not allowed. The assumption that the variables are nominal is in practice not a restriction. Many variables are nominal and other types of variables can easily be re-expressed in terms of nominal variables. As an example of the available amount of auxiliary information, the auxiliary variables of the Housing Demand Survey 1977/1978 is listed below.

(1) year of birth
(2) sex
(3) marital status
(4) size of the family
(5) structure of the family
(6) type of housing
(7) number of floors in the housing
(8) year of construction of the housing
(9) municipality
(10) quarter of town
(11) degree of urbanization

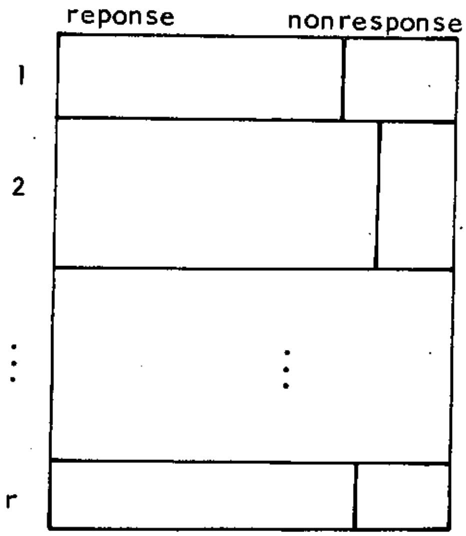
## 3.2 Graphical Methods

As a preliminary tool in the selection of auxiliary variables graphical methods have been developed. The advantage of graphical methods is clear. They bring out hidden facts and relationships and can stimulate as well as aid the analysis. They often offer a more complete and better balanced understanding then could be obtained from tabular or textual forms of presentation. Furthermore the visual relationships in the plots are more clearly grasped and more easily remembered. (See Schmid (1954).) Two simple graphical devices are presented in the next sections: the box-plot and the windmill-plot.

### 3.2.1 The box-plot

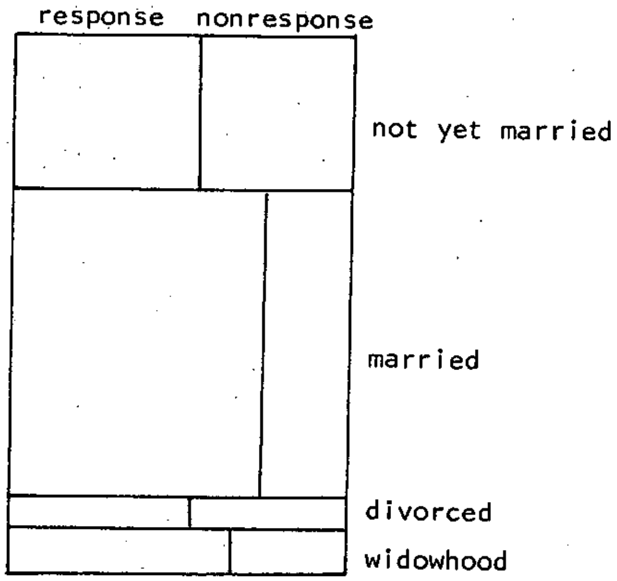
The box-plot can be seen as a generalization of a histogram or bar chart. The name of the box plot is derived from its form (see fig. 2).

FIGURE 2. THE BOX-PLOT



A rectangle of standard width and a height proportional to the sample size represents the sample. The rectangle is divided in a number of layers (the categories of the auxiliary variable). The height of a particular layer is proportional to the number of sample elements in the corresponding category. Each layer is divided by a vertical line in a left-hand part (the response) and a right-hand part (the nonresponse). The areas of these two parts are proportional to the amounts of response and nonresponse in the particular category. Fig. 3 contains an example of a box-plot. The data originate from the Housing Demand Survey 1977/1978 as far as it concerns Amsterdam. The auxiliary variable is the marital status of the person in the sample.

FIGURE 3. BOX-PLOT OF MARITAL STATUS IN AMSTERDAM IN THE HOUSING DEMAND SURVEY 1977/1978.



A number of aspects may be worth paying attention to:

(1) The heights of the layers indicate to what extent categories contribute to the sample. Clearly a large part of the people is married. The smallest category is the category of people who are divorced.

(2) The extent of the nonresponse can be read from the distance of the vertical dividing lines to the right-hand side of the box. In this example there obviously is a considerable amount of nonresponse.

(3) If all dividing lines form approximately a straight line there is no relationship between response behaviour and the auxiliary

variable. Clearly, in this situation there exists a relationship: Married people respond better than other people. Response is bad in the group of unmarried and divorced people.

More about the box plot can be found in Bethlehm & Kersten (1981).

### 3.2.2 The Windmill-Plot

The windmill-plot is a graphical representation of the results of correspondence analysis. Correspondence analysis is a technique for the analysis of associations in two-way tables. (See e.g. Benzecri (1976).). A geometrical representation of the rows (the categories of the vertically tabulated variable) and the columns (the categories of the horizontally tabulated variable) is constructed. This geometrical representation contains all the information concerning the associations in the table. By means of a scaling procedure rows and columns are assigned values in such a way that the correlation coefficient, computed by using these values, is maximized. To each cell in the table there correspond two scale values: a row-value and a column-value. When these values are conceived as coordinates, a plot of the table can be constructed. In this plot all points form an unequally spaced grid. Such a plot may not be easy to interpret. To simplify interpretation regression lines are plotted instead of the points themselves. Due to the special properties of the scale values the regression line to explain y-values from the x-values in the plot has the simple form

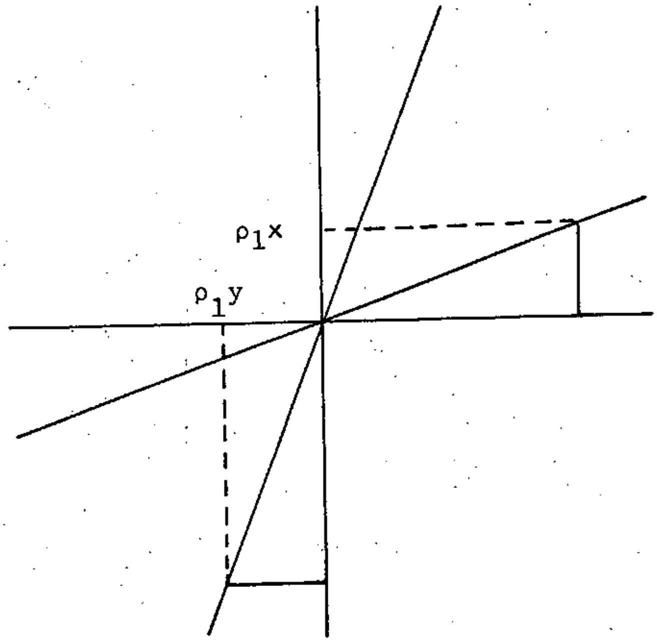
$$y = \rho_1 x \qquad (1)$$

and the regression line to explain the x-values from the y-values has the form

$$x = \rho_1 t \qquad (2)$$

were $\rho_1$ is the maximized correlation coefficient. By plotting both regression lines the result is the windmill-plot, see fig. 4.

FIGURE 4. THE WINDMILL-PLOT



A number of aspects may be worth noting:

(1) The origin represents both marginal distributions of the table

(2) Scale values close to the origin point at categories which resemble the marginal distribution and thus have a regular behaviour. Far out scale values indicate differently behaving categories.

(3) The relationship between the two variables is strong if the two regression lines are near the $45^{o}$ -line.

(4) Projection of a differently behaving category of one variable via the regression line on the axis of the other variable provides a clue about the dependencies of the categories of the variables.

The plot as described above can not account for all the information in the table. It explains as much as is possible in a two-dimensional plot. Conditionally on the first plot a second plot can be constructed, which accounts for as much as is possible of the information not yet explained. If necessary even more plots can be constructed, but preferably one plot is sufficient to explain the major part of the associations.

A total of s of such plots can be made, in which s is one less than the minimum of the number of rows and the number of columns. Let $\rho_1, \rho_2, .., \rho_s$ be the maximized correlation coefficients. Since

$$\sum_{i=1}^{s} \rho_i = X^2/N, \qquad (3)$$

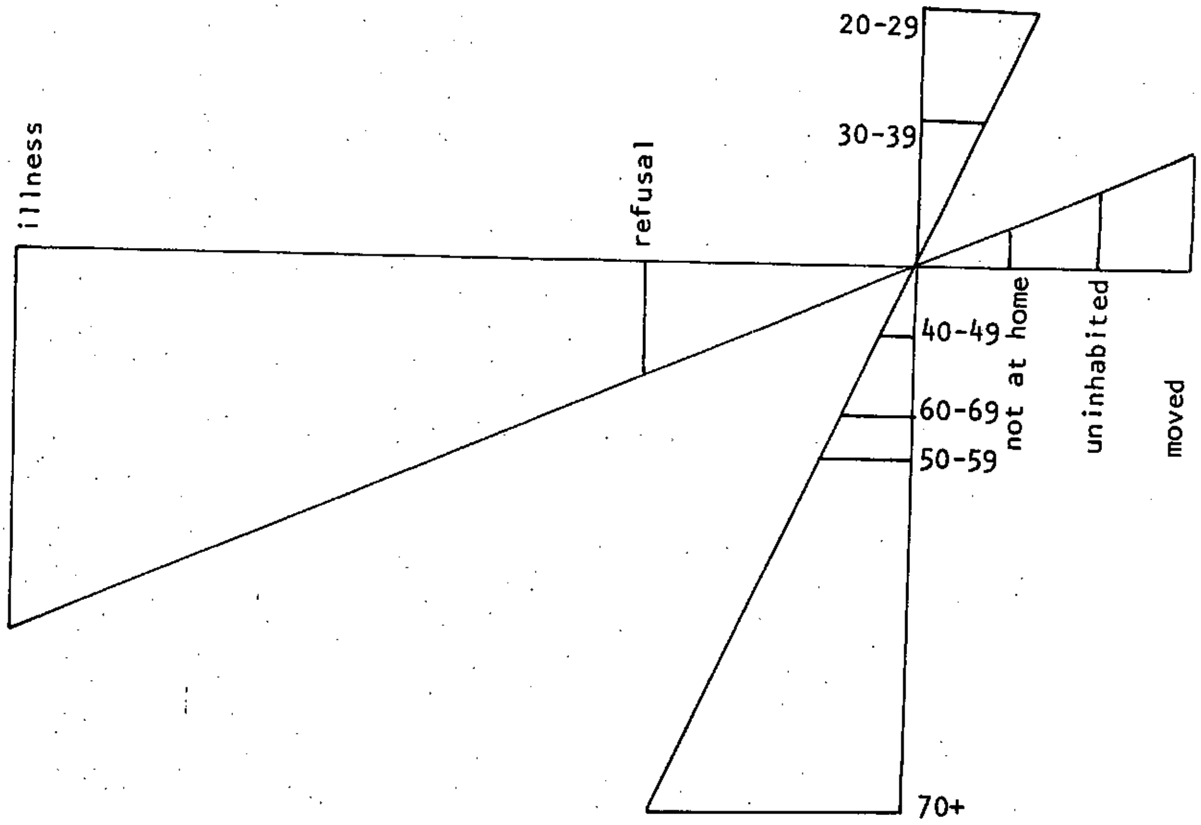
where $X^2$ is the chi-square test statistics for the table and N the general total,

$$\tau_i = N\rho_i^2/X^2 \qquad (4)$$

is a measure of the amount of information explained by the i-th plot (i=1, 2, .., s).

Fig. 5 contains the first windmill-plot for the variables age (six categories) and type of nonresponse (five categories) of the Housing Demand Survey 1977/1978 as far as it concerns Amsterdam.

FIGURE 5: WINDMILL-PLOT OF AGE BY TYPE OF NONRESPONSE IN AMSTERDAM IN THE HOUSING DEMAND SURVEY 1977/78



It contains about 88% of the information about associations in the table ($\tau_1$ = 0.88). The main reasons for nonresponse of the old people are refusal and illness. In case of young people the nonresponse is the result of the impossibility of making contact: uninhabited, not at home and moved. More about the application of correspondence analysis can be fcund in Bethlehem & Kersten (1980).

### 3.3 Other selection methods

There are many other, mainly nongraphical, method to determine the association between auxiliary vairables and the response behaviour. Much about association in contingency tables can e.g. be found in Bishop, Fienberg & Holland (1975).

A popular method for the selection of the most important auxiliary variables is AID (Automatic Interaction Detection), described by Morgan & Sonquist (1963). In a stepwise process those auxiliary variables are determined which can explain as much as possible of the variance of the binary response variable. There are disadvantages which make reliable application of this method doubtful. As the selection process proceeds in a stepwise fashion there is no guarantee that the optimal solution will be found. Because there is no stopping rule based on a statistical model this sense the result is rather arbitrary. Further research in this field is necessary (see e.g. Kass (1980)).

## 4. REDUCTION OF NONRESPONSE BIAS BY SUBGROUP WEIGHTING

When a relationship is found or suspected between the variable under investigation (Y) and the response behaviour (R) measures have to be taken in order to reduce the nonresponse bias. In this section a number of adjustment procedures are discussed which are based on subgroup weighting. Attention is focussed on estimating the population mean of Y.

It can be shown that the bias, introduced by only using response values, is proportional to the covariance between Y and R. If it would be possible to divide the population in a number of subgroups in each of which the covariance is neglectable, then (nearly unbiased) estimates of the subgroup means can be combined into a (nearly unbiased) estimate of the population mean.

Let the finite population consist of N elements $U_1$, $U_2$, .., $U_N$ with Y-values $Y_1$, $Y_2$, .., $Y_N$. From this population a simple random sample $\underline{u}_1$, $\underline{u}_2$, ..,$\underline{u}_n$

(stochastic variables are underlined) of size n is selected without replacement. The corresponding y-values are $\underline{y}_1, \underline{y}_2, .., \underline{y}_n$ and the response behaviour is indicated by $\underline{r}_1, \underline{r}_2, ..,$ ($\underline{r}_i = 1$ indicating response and $\underline{r}_i = 0$ nonresponse). In fact $\underline{y}_i$ can only be observed for those sample elements $\underline{u}_i$ for which $\underline{r}_i = 1$. The $\underline{m}$ responding elements are denoted by $u_{-1}, \underline{u}_2^*, ...., \underline{u}_{\underline{m}}^*$ ($\underline{m} = \underline{r}_1 + \underline{r}_2 + .. + \underline{r}_n$), with y-values $\underline{y}_1^*, \underline{y}_2^*, .., \underline{y}^*_{\underline{m}}$.

Let X be an auxiliary variable inducing a division of the population in H subgroups with sizes $N_1, N_2, .., N_H$. In subgroup weighting first of all in each subgroup h an estimator $\bar{\underline{y}}_h^*$ for the subgroup mean is computed:

$$\bar{\underline{y}}_h^* = \frac{1}{\underline{m}_h} \sum_{i=1}^{\underline{m}_h} \underline{y}_{hi}^*, \qquad (h = 1, 2, .., H) \tag{5}$$

where $\underline{y}_{h1}^*, \underline{y}_{h2}^*, ...., \underline{y}^*_{h\underline{m}_h}$ are the values of the $\underline{m}_h$ responding elements in subgroup h. The subgroup estimators $\bar{\underline{y}}_1^*, \bar{\underline{y}}_2^*, .., \bar{\underline{y}}_H^*$ are combined into a population estimators $\bar{\underline{y}}^*$.

$$\bar{\underline{y}}^* = \sum_{h=1}^{H} \underline{w}_h \bar{\underline{y}}_h^* \tag{6}$$

The type of estimator is determined by the available amount of information about the weights $\underline{w}_1, \underline{w}_2, .., \underline{w}_H$.

If the sizes $N_1, N_2, .., N_H$ of the subgroups are known the situation is equivalent to poststratification. (See e.g. Holt & Smith (1979).) The weights are not random but fixed quantities:

$$\underline{w}_h = \frac{N_h}{N} \qquad (h = 1, 2, .., H) \tag{7}$$

If these sizes are not known they can be estimated by

$$\underline{w}_h = \frac{\underline{n}_h}{n}, \qquad (h = 1,2,..,H) \qquad (8)$$

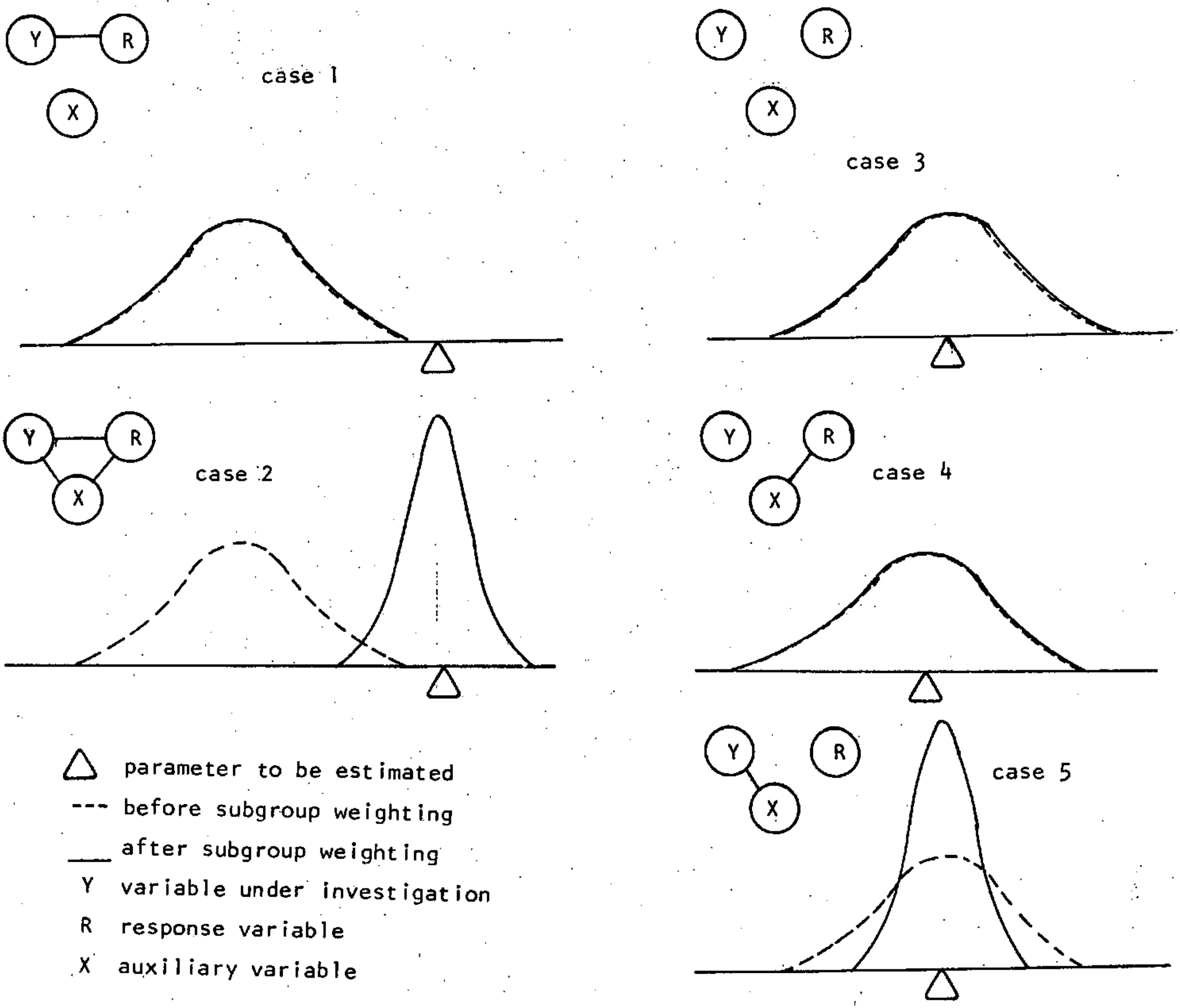
where $\underline{n}_h$ is the number of sample elements in subgroup h $(n=\underline{n}_1 + \underline{n}_2 + .. + \underline{n}_H)$.

In an intermediate situation where two auxiliary variables $X_1$ and $X_2$ are used and only the marginal totals of the two variables are known, a raking procedure can be applied to estimate the weights (see e.g. Chapman (1976)). Suppose $X_1$ induces G groups and $X_2$ induces H groups. Crossing $X_1$ and $X_2$ results in a subdivision into G x H groups. If only the marginal totals $N_{g+}$ $(g=1,2,..,G)$ of $X_1$ and $N_{+h}$ $(h=1,2,..,H)$ of $X_2$ are known then by using the sample information good estimates $\underline{N}_{gh}$ of $N_{gh}$ can be computed. The weights are then equal to

$$\underline{w}_{gh} = \frac{\underline{N}_{gh}}{N} \qquad (g=1,2,..,G;\ h=1,2,..,H) \qquad (9)$$

All three estimators have, when used in the same grouping situation, the same bias, but the greater the amount of available information on the subgroup sizes the smaller the variance of the estimate. Subgroup weighting has two advantages: reduction of the variance of the estimate and reduction of the response bias. The most extreme possibilities are illustrated in fig. 6. If two variables are connected it means that they have a strong correlation.

FIG. 6. VARIANCE AND BIAS OF ESTIMATORS BEFORE AND AFTER SUBGROUP WEIGHTING


Y
R
X
case 1
Y
R
X
case 3
Y
R
X
case 2
Y
R
X
case 4
Y
R
X
case 5
△ parameter to be estimated
--- before subgroup weighting
___ after subgroup weighting
Y variable under investigation
R response variable
X auxiliary variable

A number of conclusions can be drawn:

(1) If nonresponse bias exists subgroup weighting is significant when X and R are correlated (case 2). Both bias and variance are reduced.

(2) If no nonresponse bias exists a correlation between X and R has no effect (case 4). Only correlation between X and Y reduces the variance (case 5).

Because the data on the nonrespondents are missing, it is impossible to use the remaining data to find an auxiliary variable X which is highly correlated with Y. It is, however, possible to use this data to look for auxiliary variables which are highly correlated with the response variable R. If such a variable has been found, application of it in subgroup weighting will reduce the nonresponse bias (if it exists), but not always the variance.

## 5. Other adjustment methods

Several other adjustment methods appear in the literature. Several of them will be discussed in this section. Some of them need further research to establish their merits.

### 5.1 No adjustment

In some situations no adjustment is necessary. If it appears that no relationship exists between the variable under investigation and the response behaviour the response can be considered as a random sample from the population. Also if statements are restricted to the population of potential respondents no correction is necessary. In all other situations no adjustment is only justified if the category "nonresponse" is included in all tables in publications.

## 5.2. Imputation

Imputation procedures solve the problem of missing observations due to nonresponse by substitution of values in the records of the nonrespondents. In "hot deck" imputation data are taken from respondents of the current survey, while in "cold deck" imputation data are taken from a previous survey. If the response structure of previous and current survey resemble each other the results of cold deck imputation and hot deck imputation will roughly be the same. Imputation can be carried out in several ways. Some of them are:

(1) imputation of a random respondent
(2) imputation of the mean respondent
(3) imputation of a random respondent within the same subgroup
(4) imputation of the mean respondent within the same subgroup
(5) imputation of a value obtained by fitting a model
(6) imputation of upper or lower bounds

Procedures (1) and (2) do not reduce the bias. Procedures (3) and (4) resemble subgroup weighting. The effect of procedure (5) depends strongly on the fit of the model and the reasonableness of the model assumptions. Procedure (6) gives insight in how bad things could be if no adjustment would take place.

## 5.3. Adjustment for not-at-homes

The well-known method of Politz & Simmons (1949) tries to adjust for not-at-home bias by estimating the probability to find a person at home. This is performed by asking respondents e.g. how often they were at home at the time of the interview during the previous days. The at-home-probability, constructed in this way, can be used as a stratification variable. It is also worth trying to find a model which explains the relationship between the variable under investigation and the at-home-probability. Extrapolation of this model to the group of not-at-homes may provide more information about this group.

## 5.4. Adjustment for refusers

It is possible to measure the willingness of people to co-operate in the survey (see Van Tulder (1977)). Using this information a procedure analogous to adjustment for not-at-homes can be carried out. Furthermore the willingness to co-operate is a measure for the survey climate. The construction of a scale to obtain this information will probably be somewhat more difficult then in the case of not-at-home adjustment.

## 5.5 Double sampling

In order to get more information about nonrespondents Hansen & Hurwitz (1946) propose selecting a sample from the nonrespondents. Specially trained interviewers try as yet to obtain (part of) the missing information. Time and money constraints often prevent application of double sampling.

## 5.6. The principal question

If the method of Hansen & Hurwitz is too expensive the principal question procedure may offer a substitute. In many surveys there often is one important basic question around which the survey has been constructed. If during the field work problems are met with completing the whole questionnaire, the interviewer may try to get an answer on only the principal question. This may even be tried afterwards by letter or by telephone. This technique will shortly be tried out in one of the surveys of the CBS.

## 6. Conclusions

In view of the rise in nonresponse rates during the past years it is important to carry out thorough research on the impact of nonresponse on the quality of the survey.

Quite a few adjustment procedures appear in literature, which all aim at reduction of the nonresponse bias. A comparative study of these procedures has to provide decisive answers about their merits.

The large differences which exist with regard to objective, design and execution of surveys prevent correct interpretation of differences in nonresponse figures. It is therefore necessary to create a theoretical framework which allows proper comparison.

Of course reduction of nonresponse during the field work will remain an important topic.

## REFERENCES

[1] Benzécri, J.P. (1976) L'Analyse des Données. Dunod, Paris.

[2] Bethlehem J.G. and Kersten, H.M.P. (1981), "Graphical Methods in Non-Response Analysis and Sample Estimation", Staatsuitjeverij, The Hague.

[3] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975), Discrete Multivariate Analysis, MIT Press, Cambridge.

[4] Chapman, D.W. (1976), "A Survey of Non-Response Imputation Procedures", of the American Statistical Association, Social Statistics Section, 245-251.

[5] Hansen, M.H. and Hurwitz, W.N. (1946), "The Problem of Non-Response in Sample Surveys", Journal of the American Statistician, 41, 517-529.

[6] Holt, D. and Smith, I.M.F. (1979), "Post Stratification", Journal of the Royal Statistical Society, series A, 142, 33-46.

[7] Kass, G.V. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorial Data", Applied Statistics, 29, 199-217.

[8] Kish, L. (1967), Survey Sampling, Wiley, New York.

[9] Morgan, J.N. and Sonquist, J.A. (1963), "Problems in the Analysis of Survey Data", Journal of the American Statistical Association, 58, 415-434.

[10] Politz, A. and Simmons, W. (1949), "An Attempt To Get the Not-At-Homes into the Sample Without Callbacks", Journal of the American Statistical Association, 44, 9-31.

[11] Schmid, C.F. (1954), Handbook of Graphical Presentation, Ronald Press, New York.

[12] Tulder, J.J.M. van (1977), "Op de grens van non-response", Jaarboek van de Nederlandse Vereniging van Marktonderzoekers, 1977, 43-52.

[13] Widdershoven, M. and Berg, J. van den (1980),"Non-respons bij twee 'persoons- en gezinsenquêtes'", CBS-Select 1, 357-365. Staatsuitgeverij, The Hague.

# ON THE VARIANCES OF ASYMPTOTICALLY NORMAL ESTIMATORS FROM COMPLEX SURVEYS

David A. Binder[1]

The problem of specifying and estimating the variance of estimated parameters based on complex sample designs from finite populations is considered. The results of this paper are particularly useful when the paramtere estimators cannot be defined explicitly as a function of other statistics from the sample. It is shown how these results can be applied to linear regression, logistic regression and loglinear contingency table models.

## 1. INTRODUCTION

In recent years, there has been an increasing demand for using survey data to estimate the parameters of traditional models such as regression parameters, discriminant functions, logit and probit parameters and others. However, for many such surveys, the primary objectives of the survey is the estimation of population or sub-population means, totals, trends and so on. For this reason and because of operational considerations, the survey design is often not a simple random sample, but is more typically stratified and often multi-stage with possibly unequal probabilities at certain stages of sampling.

Because of this, there has been much discussion (see, for example, Sarndal;1978) on whether the sampling weights should be used in making inferences about these model parameters. The answer seems to depend on whether a superpopulation model is appropriate for all population units. If this is the case, the inference on the superpopulation parameters is often the primary concern. This leads to model-based inference, where, for a given sample, the inferences do not depend on the sampling weights.

[1] D.A. Binder, Institutional and Agriculture Survey Methods Division, Statistics Canada.

The question that comes to mind is: If the superpopulation model is not appropriate, what parameters are we estimating? It must be recognized that for many studies, particularly in the social sciences, the model (e.g. linear regression) is only a convenient approximation of the real world and the parameters of that model (e.g. correlations and partial correlations) are often used to understand the approximate interdependencies of the variables rather than having a particular scientific interpretation. Therefore, the parameters we are estimating do not necessarily refer to a true superpopulation model, but are of a more descriptive nature.

In this paper, we adopt the view that we are interested in making inferences about these "descriptive" parameters of the population. For example, suppose $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ are N × p and N × 1 matrices respectively, where each row of $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ corresponds to a different individual of the population. We are interested in the descriptive parameter, $\underset{\sim}{B}$, a p×1 vector satisfying the equations:

$$\underset{\sim}{X}^T \underset{\sim}{X} \underset{\sim}{B} = \underset{\sim}{X}^T \underset{\sim}{Y} \qquad (1.1)$$

This view of descriptive parameters is the same as that taken by Frankel (1971) and Kish and Frankel (1974).

The usual estimation of such parameters normally takes into account the sampling weights. If we denote by $\pi_i$ the probability that the i-th unit in the sample is sampled and let $\underset{\sim}{\Pi} = \text{diag}\ (\pi_1, \ldots, \pi_n)$, then the weighted parameter estimate for $\underset{\sim}{B}$ satisfies:

$$\underset{\sim}{x}^T \underset{\sim}{\Pi}^{-1} \underset{\sim}{x} \underset{\sim}{B} = \underset{\sim}{x}^T \underset{\sim}{\Pi}^{-1} \underset{\sim}{y}, \qquad (1.2)$$

where $\underset{\sim}{x}$ and $\underset{\sim}{y}$ are n×p and n×1 matrices respectively, the rows of which correspond to the sampled rows of $\underset{\sim}{X}$ and $\underset{\sim}{Y}$.

Suppose, now, an estimator of a population parameter can be expressed as:

$$\hat{\theta} = g(z_1, \ldots, z_k), \qquad (1.3)$$

where $E(z_i) = Z_i$. Here, $\hat{\theta}$ is an estimator of $g(Z_1, \ldots, Z_k)$. Following Tepping (1968) and Woodruff (1971), a Taylor series expansion for $\hat{\theta}$ yields:

$$V[\hat{\theta}] \doteq V[\sum_{i=1}^{k} (\frac{\partial g}{\partial Z_i})(z_i - Z_i)] \ . \tag{1.4}$$

These formulae are exemplified for estimation of regression coefficients (1.1) by Tepping (1968). However, the expressions resulting from (1.4) for the variances of the regression coefficients are somewhat complicated compared to those derived by Fuller (1975).

In this paper we consider parameters which are not defined through an explicit equation such as (1.3), but instead are defined implicitly as $\underset{\sim}{U}(\underset{\sim}{Z},\underset{\sim}{\theta}) = 0$. A simple example showing the distinction would be the ratio parameter:

$$R = \frac{\Sigma Y_k}{\Sigma X_k} ,$$

which could also be defined implicitly as:

$$\Sigma Y_k - R\Sigma X_k = 0.$$

When we deal with some models such as indirect loglinear models or logistic regression models, the parameters can be defined only through implicit relationships. The extension of Tepping's (1968) results for this case is fairly straightforward,but does not appear in its general form at present in the literature. There are, however, specific examples of its application; see, for example Fuller (1975) and Freeman and Koch (1976).

In Section 2 we give the general framework and the main results of the paper. A number of models are exemplified in Section 3.

## 2. GENERAL FRAMEWORK

### 2.1 Framework

The population units are labelled 1, ..., N. Associated with the i-th unit we have a q-dimensional data vector $\underset{\sim}{X}_i$. We have a parameter space $\underline{\Theta} \subseteq R^P$. The parameter $\underset{\sim}{\theta}_o = (\theta_{1o}, \ldots, \theta_{po})$ is defined by the p equations:

$$U_i(\underset{\sim}{X},\underset{\sim}{\theta}_o) = \sum_{k=1}^{N} u_i(\underset{\sim}{X}_k,\underset{\sim}{\theta}_o) - v_i(\underset{\sim}{\theta}_o) = 0, \tag{2.1}$$

for i=1, ..., p. We assume that equations (2.1) define $\underset{\sim}{\theta}_o$ uniquely in $\Theta$. We also assume that $\partial u_i(\underset{\sim}{X},\underset{\sim}{\theta})/\partial\underset{\sim}{\theta}$ and $\partial v_i(\underset{\sim}{\theta})/\partial\underset{\sim}{\theta}$ exist in a neighbourhood of $\underset{\sim}{\theta}_o$. A simple example of (2.1) is where $\theta_o$ is a population total, and we have $U(\underset{\sim}{X},\theta_o) = \sum_{k=1}^{N} X_k - \theta_o$. Here, $u(X_k,\theta_o) = X_k$ and $v(\theta_o) = \theta_o$.

We select a sample of the units, according to some probability distribution defined on the set of all non-empty subsets of {1, ..., N}. We denote by $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ the selected values of $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_N$. We assume that for any $\underset{\sim}{\theta} \in \Theta$, we can construct a consistent, asymptotically normal estimator of $U_i(\underset{\sim}{X},\underset{\sim}{\theta})$. We denote this estimator by $\hat{U}_i(\underset{\sim}{x},\underset{\sim}{\theta})$. For example, for many without replacement sampling schemes,

$$\hat{U}_i(\underset{\sim}{x},\underset{\sim}{\theta}) = \sum_{k=1}^{n} u_i(\underset{\sim}{x}_k,\underset{\sim}{\theta})/\pi_k - v_i(\underset{\sim}{\theta}) \tag{2.2}$$

will be a consistent asymptotically normal estimator, where $\pi_k$ is the probability of inclusion for the k-th unit.

We let $\sigma_{ij}(\underset{\sim}{X},\underset{\sim}{\theta}) = \mathrm{Cov}[\hat{U}_i(\underset{\sim}{x},\underset{\sim}{\theta}), \hat{U}_j(\underset{\sim}{x},\underset{\sim}{\theta})]$. For example, for estimator (2.2), we have:

$$\sigma_{ij}(\underset{\sim}{X},\underset{\sim}{\theta}) = \sum_{k=1}^{N} \sum_{\ell=1}^{N} u_i(\underset{\sim}{X}_k,\underset{\sim}{\theta})\, u_j(\underset{\sim}{X}_\ell,\underset{\sim}{\theta})(\pi_{k\ell}-\pi_k\pi_\ell)/\pi_k\pi_\ell, \tag{2.3}$$

where $\pi_{k\ell}$ is the probability that the k-th and ℓ-th units in sample.

We let $\underset{\sim}{\Sigma}(\underset{\sim}{X},\underset{\sim}{\theta})$ be the p×p matrix with entries $\sigma_{ij}(\underset{\sim}{X},\underset{\sim}{\theta})$, and $\hat{\underset{\sim}{\Sigma}}(\underset{\sim}{x},\underset{\sim}{\theta})$ be a consistent estimator for $\underset{\sim}{\Sigma}$. Now, for any given $\underset{\sim}{\theta}$,

$$U_i(\underset{\sim}{X},\underset{\sim}{\theta}) + v_i(\underset{\sim}{\theta}) = \sum_{k=1}^{N} u_i(\underset{\sim}{X}_k,\underset{\sim}{\theta}),$$

so that estimators $\hat{U}_i(\underset{\sim}{X},\underset{\sim}{\theta})$ and $\hat{\underset{\sim}{\Sigma}}(\underset{\sim}{x},\underset{\sim}{\theta})$ can be specified for any design in which we can derive consistent asymptotically normal estimators of population totals and consistent estimators for the variances of the estimators of the totals.

The Horvitz-Thompson estimator for (2.3) is:

$$\sum_{k=1}^{n} \sum_{\ell=1}^{n} u_i(\underset{\sim}{x}_k,\underset{\sim}{\theta}) \quad u_j(\underset{\sim}{x}_\ell,\underset{\sim}{\theta})(\pi_{k\ell} - \pi_k\pi_\ell)/\pi_k\pi_\ell\pi_{k\ell} \, . \tag{2.4}$$

In the case of fixed sample size, the Yates-Grundy estimator of (2.3) is:

$$\sum_{k<\ell}\sum \left[\frac{u_i(\underset{\sim}{x}_k,\underset{\sim}{\theta})}{\pi_k} - \frac{u_i(\underset{\sim}{x}_\ell,\underset{\sim}{\theta})}{\pi_\ell}\right]\left[\frac{u_j(\underset{\sim}{x}_k,\underset{\sim}{\theta})}{\pi_k} - \frac{u_j(\underset{\sim}{x}_\ell,\underset{\sim}{\theta})}{\pi_\ell}\right] (\pi_k\pi_\ell - \pi_{k\ell}). \tag{2.5}$$

Letting $\underset{\sim}{U}(\underset{\sim}{X},\underset{\sim}{\theta})$ and $\hat{\underset{\sim}{U}}(\underset{\sim}{x},\underset{\sim}{\theta})$ be the p-dimensional vectors with components $U_i(\underset{\sim}{X},\underset{\sim}{\theta})$ and $\hat{U}_i(\underset{\sim}{x},\underset{\sim}{\theta})$ respectively, we define

$$\underset{\sim}{J}(\underset{\sim}{X},\underset{\sim}{\theta}) = \partial\underset{\sim}{U}(\underset{\sim}{X},\underset{\sim}{\theta})/\partial\underset{\sim}{\theta} \tag{2.6}$$

$$\hat{\underset{\sim}{J}}(\underset{\sim}{x},\underset{\sim}{\theta}) = \partial\hat{\underset{\sim}{U}}(\underset{\sim}{x},\underset{\sim}{\theta})/\partial\underset{\sim}{\theta}, \tag{2.7}$$

where $\underset{\sim}{J}$ and $\hat{\underset{\sim}{J}}$ are p×p partial derivative matrices. Assume that the matrices are continuous functions of $\underset{\sim}{\theta}$ and that the partial derivatives with respect to $\underset{\sim}{\theta}$ exist in a neighbourhood of $\underset{\sim}{\theta}_o$. Also assume $\hat{\underset{\sim}{J}}(\underset{\sim}{x},\underset{\sim}{\theta})$ is a consistent estimator of $\underset{\sim}{J}(\underset{\sim}{X},\underset{\sim}{\theta})$.

Our estimator for $\underset{\sim}{\theta}$ is given by $\hat{\underset{\sim}{\theta}}$, the solution to:

$$\hat{U}_i(\underset{\sim}{x},\hat{\underset{\sim}{\theta}}) = 0, \text{ for } i=1, \ldots, p. \tag{2.8}$$

We assume the sample size is sufficiently large so that the solution to (2.8) is unique in $\Theta$. We show in the next section that the covariance matrix of $\underset{\sim}{\hat{\theta}}$ can be consistently estimated by:

$$[\underset{\sim}{\hat{J}}^{-1}(\underset{\sim}{x},\underset{\sim}{\hat{\theta}})]\ \underset{\sim}{\hat{\Sigma}}(\underset{\sim}{x},\underset{\sim}{\hat{\theta}})\ [\underset{\sim}{\hat{J}}^{-1}(\underset{\sim}{x},\underset{\sim}{\hat{\theta}})]^T.$$

## 2.2 Asymptotic Theory

Following the asymptotic arguments of Madow (1948), and Hájek (1960), we consider a sequence of populations indexed by t, with sizes $N^{(t)}$ and data $\underset{\sim}{X}^{(t)}$. We assume $N^{(t)} \to \infty$ as $t \to \infty$. For population t, we select a sample of size $n^{(t)}$ and observe data $\underset{\sim}{x}^{(t)}$. We let $\nu^{(t)} = E(n^{(t)})$ and assume

$$\lim_{t\to\infty} \nu^{(t)} = \infty$$

$$\lim_{t\to\infty} (N^{(t)} - \nu^{(t)}) = \infty$$

For any $\underset{\sim}{\theta}$ in a neighbourhood of $\underset{\sim}{\theta}_o^{(t)}$ we assume

$$[\nu^{(t)}]^{\frac{1}{2}}\ [\underset{\sim}{\hat{U}}(\underset{\sim}{x}^{(t)},\ \underset{\sim}{\theta}) - \underset{\sim}{U}(\underset{\sim}{X}^{(t)},\underset{\sim}{\theta})]/N^{(t)}$$

is asymptotically $N[0,\underset{\sim}{S}(\underset{\sim}{\theta})]$, where

$$\underset{\sim}{S}(\underset{\sim}{\theta}) = \lim[\nu^{(t)}\ \underset{\sim}{\Sigma}(\underset{\sim}{X}^{(t)},\underset{\sim}{\theta})/\{N^{(t)}\}^2]$$

exists. We assume

$\underset{\sim}{K}(\underset{\sim}{\theta}) = \lim\ \underset{\sim}{J}(\underset{\sim}{X}^{(t)},\underset{\sim}{\theta})/N^{(t)}$ exists and also

$\text{plim}\ \underset{\sim}{\hat{J}}(\underset{\sim}{x}^{(t)},\underset{\sim}{\theta})/N^{(t)} = \underset{\sim}{K}(\underset{\sim}{\theta})$.

Also, we assume

$$\lim[\text{rank}\ \{\underset{\sim}{J}(\underset{\sim}{X}^{(t)},\underset{\sim}{\theta})\}] = \text{plim}[\text{rank}\ \{\underset{\sim}{\hat{J}}(\underset{\sim}{x}^{(t)},\underset{\sim}{\theta})\}] = p.$$

We define $\underset{\sim}{\hat{\theta}}^{(t)}$ to satisfy

$$\underset{\sim}{\hat{U}}(\underset{\sim}{x}^{(t)},\ \underset{\sim}{\hat{\theta}}^{(t)}) = 0.$$

By a Taylor series expansion, we obtain

$$\underset{\sim}{\hat{U}}(\underset{\sim}{x}^{(t)},\ \underset{\sim}{\theta}_o^{(t)}) \doteq -\ \underset{\sim}{\hat{J}}(\underset{\sim}{x}^{(t)},\ \underset{\sim}{\hat{\theta}}^{(t)})\ (\underset{\sim}{\hat{\theta}}^{(t)} - \underset{\sim}{\theta}_o^{(t)}). \qquad (2.9)$$

Since the left hand side of (2.9) is asymptotically normal, we have that

$$(n^{(t)})^{\frac{1}{2}}\ (\underset{\sim}{\hat{\theta}}^{(t)} - \underset{\sim}{\theta}_o^{(t)})$$

is asymptotically $N[\underset{\sim}{0},\ \underset{\sim}{G}(\underset{\sim}{\theta}_o)]$, where $\underset{\sim}{S}(\underset{\sim}{\theta}_o) = \underset{\sim}{K}(\underset{\sim}{\theta}_o)\ \underset{\sim}{G}(\underset{\sim}{\theta}_o)[\underset{\sim}{K}(\underset{\sim}{\theta}_o)]^T$.

Therefore,

$$\underset{\sim}{G}(\underset{\sim}{\theta}_o) = [\underset{\sim}{K}^{-1}(\underset{\sim}{\theta}_o)]\ \underset{\sim}{S}(\underset{\sim}{\theta}_o)\ [\underset{\sim}{K}^{-1}(\underset{\sim}{\theta}_o)]^T \qquad (2.10)$$

and a consistent estimator for $\underset{\sim}{G}(\underset{\sim}{\theta}_o)$ is :

$$n^{(t)}[\hat{\underset{\sim}{J}}^{-1}(\underset{\sim}{x},\hat{\underset{\sim}{\theta}})]\ \hat{\underset{\sim}{\Sigma}}(\underset{\sim}{x},\hat{\underset{\sim}{\theta}})\ [\hat{\underset{\sim}{J}}^{-1}(\underset{\sim}{x},\hat{\underset{\sim}{\theta}})]^T. \qquad (2.11)$$

Hence, when the functional form of $\hat{\underset{\sim}{U}}(\underset{\sim}{x},\underset{\sim}{\theta})$ and $\hat{\underset{\sim}{\Sigma}}(\underset{\sim}{x},\underset{\sim}{\theta})$ is specified, we need only derive the matrix $\hat{\underset{\sim}{J}}(\underset{\sim}{X},\underset{\sim}{\theta}_o)$ and its estimator $\hat{\underset{\sim}{J}}(\underset{\sim}{x},\hat{\underset{\sim}{\theta}})$ to use these results.

## 3. EXAMPLES

### 3.1 Introduction

In this section we consider in detail the implication of the general formulation given in Section 2 with respect to estimating the variances of certain population parameter estimators. In particular, we discuss ratios, regression coefficients and log linear models for categorical data. Other models, such as probit models could be analyzed analogously.

In general, we use the following notation. If $\underset{\sim}{W}_1, \ldots, \underset{\sim}{W}_N$ are population values, with $\underset{\sim}{W} = \Sigma\underset{\sim}{W}_k$, then on selecting a sample $\underset{\sim}{w}_1, \ldots, \underset{\sim}{w}_n$, we have an unbiased estimator of $\underset{\sim}{W}$ given by $\hat{\underset{\sim}{W}}$. We let $\underset{\sim}{V}(\hat{\underset{\sim}{W}})$ represent the covariance matrix for $\hat{\underset{\sim}{W}}$ and $\hat{\underset{\sim}{V}}(\hat{\underset{\sim}{W}})$ a consistent estimator of $\underset{\sim}{V}(\hat{\underset{\sim}{W}})$. The particular form of this estimator will depend on the sample design; for example, multi-stage stratified, pps with replacement, etc. .

### 3.2 Ratios

Suppose we are interested in $R = \Sigma X_{k2}/\Sigma X_{k1}$. We define

$$U(\underset{\sim}{X},R) = \Sigma X_{k2} - R\Sigma X_{k1}.$$

Therefore, for without replacement sampling, we have :

$$\hat{U}(\underset{\sim}{x},R) = \hat{X}_2 - R\hat{X}_1.$$

Setting $\hat{U}(\underset{\sim}{x},\hat{R}) = 0$, we obtain

$$\hat{R} = \hat{X}_2/\hat{X}_1 . \tag{3.1}$$

We define $W_k = X_{k2} - R\ X_{k1}$.

Since, $J(\underset{\sim}{X},R) = -\Sigma X_{k1}$, we have that $V(\hat{R})$ is approximately $V(\hat{W})/(\Sigma X_{k1})^2$. This is estimated by $\hat{V}(\hat{W})/\hat{X}_1^2$. In the case of stratified sampling, this yields the same result as in Woodruff (1971).

## 3.3 Regression Coefficients and R

Suppose our data matrix $\underset{\sim}{X}$ is partitioned into $[\underset{\sim}{Z}|\underset{\sim}{Y}]$, the first column of $\underset{\sim}{Z}$ being the vector of 1's. The vector $\underset{\sim}{Y}$ is N×1. We have parameters of interest $\theta$, $\underset{\sim}{B}$, and $R^2$ defined by:

$$U_1 = \theta - \underset{\sim}{Y}^T \underset{\sim}{1} = 0, \tag{3.2a}$$

$$\underset{\sim}{U}_2 = \underset{\sim}{Z}^T \underset{\sim}{Z}\ \underset{\sim}{B} - \underset{\sim}{Z}^T \underset{\sim}{Y} = 0\ , \tag{3.2b}$$

$$U_3 = (\underset{\sim}{Y}^T\underset{\sim}{Y}-N^{-1}\theta^2)(R^2-1) + \underset{\sim}{Y}^T \underset{\sim}{Y} - \underset{\sim}{Y}^T \underset{\sim}{Z}\ \underset{\sim}{B} = 0. \tag{3.2c}$$

Here, $\underset{\sim}{B}$ denotes the vector of regression coefficients, $R^2$ is the coefficient of multiple determination and $\theta$ is the total of the Y's . We first consider the case where N is known. We let $SSY = \underset{\sim}{Y}^T\underset{\sim}{Y} - N^{-1}\ \theta^2$. We also define $\underset{\sim}{S}_{ZZ}$ as the estimator for $\underset{\sim}{Z}^T\underset{\sim}{Z}$, $S_{YY}$ the estimator for $\underset{\sim}{Y}^T\underset{\sim}{Y}$ and $\underset{\sim}{S}_{ZY}$ the estimator for $\underset{\sim}{Z}^T\underset{\sim}{Y}$. We therefore have :

$$\hat{\theta} = \hat{Y}\ , \tag{3.3a}$$

$$\hat{\underset{\sim}{B}} = \underset{\sim}{S}_{ZZ}^{-1}\ \underset{\sim}{S}_{ZY}\ , \tag{3.3b}$$

$$\hat{R}^2 = 1 - \frac{S_{YY} - \hat{\underset{\sim}{B}}^T\ \underset{\sim}{S}_{ZY}}{S_{YY} - N^{-1}\ \hat{Y}^2}\ , \tag{3.3c}$$

and

$$\underset{\sim}{J} = \partial\underset{\sim}{U}(\underset{\sim}{Z},\underset{\sim}{Y},\underset{\sim}{B},R^2,\theta)/\partial(\underset{\sim}{B},\hat{R},\theta) = \begin{bmatrix} \underset{\sim}{0}^T & 0 & 1 \\ \underset{\sim}{Z}^T\underset{\sim}{Z} & \underset{\sim}{0} & \underset{\sim}{0} \\ -\underset{\sim}{Y}^T\underset{\sim}{Z} & SSY & 2\bar{Y}(1-R^2) \end{bmatrix},$$

where $\bar{Y} = \theta/N$ .

Therefore,

$$\underset{\sim}{J}^{-1} = \begin{bmatrix} \underset{\sim}{0} & (\underset{\sim}{Z}^T\underset{\sim}{Z})^{-1} & \underset{\sim}{0} \\ -2\bar{Y}(1-R^2)/SSY & \underset{\sim}{B}^T/SSY & 1/SSY \\ 1 & \underset{\sim}{0}^T & 0 \end{bmatrix} .$$

Now, letting $\underset{\sim}{W}_k^T(B) = (Z_{k1}\ e_k, \ldots, Z_{kp}\ e_k)$, where $e_k = Y_k - \sum_j Z_{kj}\ B_j$, we obtain:

$$\underset{\sim}{V}[\hat{\underset{\sim}{B}}] \doteq (\underset{\sim}{Z}^T\underset{\sim}{Z})^{-1}\ \underset{\sim}{V}[\hat{\underset{\sim}{W}}(\underset{\sim}{B})](\underset{\sim}{Z}^T\underset{\sim}{Z})^{-1} . \qquad (3.4)$$

This is a direct consequence of (2.10). Note that the set of $\underset{\sim}{W}_k(\underset{\sim}{B})$ vectors corresponds to $\underset{\sim}{U}_2$ in (3.2b). Fuller (1975) obtains the same result for stratified or two-stage stratified sampling.

To estimate (3.4) we use:

$$\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{B}}] = \underset{\sim}{S}_{ZZ}^{-1}\ \hat{\underset{\sim}{V}}[\hat{\underset{\sim}{W}}(\hat{\underset{\sim}{B}})]\ \underset{\sim}{S}_{ZZ}^{-1}.$$

We can also estimate the variance of $\hat{R}^2$. If $\underset{\sim}{W}_k^T(B,R^2) = [Y_k, Z_{k1}\ e_k, \ldots, Z_{kp}\ e_k, Y_k(\sum_j Z_{kj}\ B_j - R^2Y_k)]$ and $\underset{\sim}{c}^T = [-2\hat{Y}(1-\hat{R}^2)/N, \hat{\underset{\sim}{B}}^T, 1]/(S_{YY}-N^{-1}\ \hat{Y}^2)$, we obtain:

$$\hat{V}[\hat{R}^2] \doteq \underset{\sim}{c}^T\ \hat{\underset{\sim}{V}}[\hat{\underset{\sim}{W}}(\hat{\underset{\sim}{B}},\hat{R}^2)]\ \underset{\sim}{c} . \qquad (3.5)$$

For the case where $N$ is unknown (e.g. the primary sampling units are geographic areas), we have the additional equation:

$$U_4 = N - \Sigma\ 1 . \qquad (3.6)$$

Adding the appropriate row and column to $\underset{\sim}{J}$ and inverting, we obtain the following results for estimating $V[\hat{R}^2]$.

We let

$$\underset{\sim}{W}_k^T(\underset{\sim}{B},R^2) = [Y_k, Z_{k1}\ e_k, \ldots, Z_{kp}\ e_k, Y_k(\sum_j Z_{kj}\ B_j - R^2Y_k), 1]$$

and

$$\underset{\sim}{c}^T = [-2\hat{Y}(1-\hat{R}^2)/\hat{N}, \hat{B}^T, 1, \hat{Y}^2(1-\hat{R}^2)/\hat{N}^2]/(S_{YY} - \hat{N}^{-1}\ \hat{Y}^2).$$

We then have $\hat{V}[\hat{R}^2]$ is given by (3.5) for these new values of $\underset{\sim}{W}_k(\underset{\sim}{B},R^2)$ and $\underset{\sim}{c}$.

## 3.4 Logistic Regression

As in the previous section, we assume the data matrix $\underset{\sim}{X}$ can be partitioned into $[\underset{\sim}{Z}|\underset{\sim}{Y}]$, but now $\underset{\sim}{Y}$ is a vector of 0's and 1's. In the traditional statistical framework, the logistic regression model for $\underset{\sim}{Y}$ conditional on $\underset{\sim}{Z}$ asserts that $Y_1, \ldots, Y_N$ are independent with $Pr(Y_k = 1) = p_k(\underset{\sim}{\beta})$, where :

$$p_k(\underset{\sim}{\beta}) = \frac{\exp(\underset{\sim}{\beta}^T \underset{\sim}{z}_k)}{1 + \exp(\underset{\sim}{\beta}^T \underset{\sim}{z}_k)} . \tag{3.7}$$

Letting $\underset{\sim}{B}$ be the maximum likelihood estimator for $\underset{\sim}{\beta}$, we have that $\underset{\sim}{B}$ satisfies

$$\underset{\sim}{U} = \underset{\sim}{Z}^T \underset{\sim}{P}(\underset{\sim}{B}) - \underset{\sim}{Z}^T \underset{\sim}{Y} = 0 , \tag{3.8}$$

where $\underset{\sim}{P}(\underset{\sim}{B})^T = [p_1(\underset{\sim}{B}), \ldots, p_N(\underset{\sim}{B})]$.
For a given finite population,we define $\underset{\sim}{B}$ as our parameter of interest.

We let $\underset{\sim}{C}(\underset{\sim}{B})$ be our estimate for $\underset{\sim}{Z}^T \underset{\sim}{P}(\underset{\sim}{B})$ and $\underset{\sim}{S}_{ZY}$ our estimate for $\underset{\sim}{Z}^T \underset{\sim}{Y}$. Therefore, $\hat{\underset{\sim}{B}}$ satisfies $\underset{\sim}{C}(\hat{\underset{\sim}{B}}) = \underset{\sim}{S}_{ZY}$. These equations must be solved iteratively in general. We also have

$$\underset{\sim}{J} = \frac{\partial \underset{\sim}{U}}{\partial \underset{\sim}{B}} .$$

The (i,j)th component of $\underset{\sim}{J}$ is $\sum_k Z_{ki} Z_{kj} p_k(\underset{\sim}{B}) [1-p_k(\underset{\sim}{B})]$. We denote the estimator of $\underset{\sim}{J}$ by $\hat{\underset{\sim}{J}}$.

To estimate the variance of $\hat{\underset{\sim}{B}}$, we let

$$\underset{\sim}{W}_k^T = (Z_{k1} \hat{e}_k, \ldots, Z_{kr} \hat{e}_k)$$

where $\hat{e}_k = p_k(\hat{\underset{\sim}{B}}) - Y_k$. The estimator for $\underset{\sim}{V}[\hat{\underset{\sim}{B}}]$ is given by :

$$\hat{\underset{\sim}{J}}^{-1} \hat{\underset{\sim}{V}}(\hat{\underset{\sim}{W}}) \hat{\underset{\sim}{J}}^{-1} .$$

### 3.5 Loglinear Models for Categorical Data

Suppose that each member of the population belongs to exactly one of $q$ distinct categories. Associated with category $i$ we have an $r\times 1$ vector $\underset{\sim}{a}_i$ such that the proportion of individuals in the i-th category is approximately

$$p_i(\underset{\sim}{\beta}) = \frac{\exp(\underset{\sim}{a}_i^T \underset{\sim}{\beta})}{\sum\limits_j \exp(\underset{\sim}{a}_j^T \underset{\sim}{\beta})} .$$

We let $\underset{\sim}{p}(\underset{\sim}{\beta})^T = [p_1(\underset{\sim}{\beta}), \ldots, p_q(\underset{\sim}{\beta})]$ and $\underset{\sim}{N}^T = (N_1, \ldots, N_q)$, where $N_i$ is the number of individuals in the i-th category. Now, if the population were generated from a multinomial distribution with probabilities $\underset{\sim}{p}(\underset{\sim}{\beta})$,the maximum likelihood estimator for $\underset{\sim}{\beta}$,given by $\underset{\sim}{B}$, satisfies:

$$\underset{\sim}{U} = \underset{\sim}{A}^T \underset{\sim}{N} - [\underset{\sim}{A}^T \underset{\sim}{p}(\underset{\sim}{B})] \underset{\sim}{1}^T \underset{\sim}{N} = 0,$$

where $\underset{\sim}{A}$ is a $q\times r$ matrix with i-th row being $\underset{\sim}{a}_i^T$ . We consider $\underset{\sim}{B}$ as our parameter of interest for any given finite population.

We let $\hat{\underset{\sim}{N}}$ be a consistent asymptotically normal estimator of $\underset{\sim}{N}$, with variance-covaraince matrix $\underset{\sim}{V}[\hat{\underset{\sim}{N}}]$ and estimated matrix $\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{N}}]$. Our estimator, $\hat{\underset{\sim}{B}}$, satisfies:

$$\underset{\sim}{A}^T \hat{\underset{\sim}{N}} - [\underset{\sim}{A}^T \underset{\sim}{p}(\hat{\underset{\sim}{B}})] \underset{\sim}{1}^T \hat{\underset{\sim}{N}} = 0. \qquad (3.9)$$

This estimator was suggested by Freeman and Koch (1976). It may be less efficient than Imrey, Koch and Stokes (1981, 1982) functional asymptotic regression methodology; however, we need not calculate all the components of $\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{N}}]$ to apply (3.9).

Let $\underset{\sim}{D}(\underset{\sim}{B})$ be diag$[\underset{\sim}{p}(\underset{\sim}{B})]$ and $\underset{\sim}{H}(\underset{\sim}{B}) = \underset{\sim}{D}(\underset{\sim}{B}) - \underset{\sim}{p}(\underset{\sim}{B})\ \underset{\sim}{p}(\underset{\sim}{B})^T$. We have:

$$\underset{\sim}{J} = \frac{\partial \underset{\sim}{U}}{\partial \underset{\sim}{B}} = - (\underset{\sim}{1}^T \underset{\sim}{N})\ \underset{\sim}{A}^T\ \underset{\sim}{H}(\underset{\sim}{B})\ \underset{\sim}{A} .$$

Therefore the asymptotic variance matrix for $\hat{\underset{\sim}{B}}$ is given by:

$$\underset{\sim}{V}[\hat{\underset{\sim}{B}}] = (\underset{\sim}{N}^T \underset{\sim}{1})^{-2} (\underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B}) \underset{\sim}{A})^{-1}$$

$$\underset{\sim}{A}^T(\underset{\sim}{I}-\underset{\sim}{p}(\underset{\sim}{B})\underset{\sim}{1}^T) \; V[\hat{N}] \; (\underset{\sim}{I}-\underset{\sim}{1}\, \underset{\sim}{p}(\underset{\sim}{B})^T) \; \underset{\sim}{A}(\underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B}) \underset{\sim}{A})^{-1} . \qquad (3.10)$$

This expression can sometimes be simplified as follows. If it can be assumed that $\underset{\sim}{N}/\underset{\sim}{N}^T\underset{\sim}{1} \doteq \underset{\sim}{p}(\underset{\sim}{B})$, then for $\hat{\underset{\sim}{\pi}} = \hat{\underset{\sim}{N}}/\hat{\underset{\sim}{N}}^T\underset{\sim}{1}$ we have:

$$\underset{\sim}{V}[\hat{\pi}] \doteq (\underset{\sim}{N}^T \underset{\sim}{1})^{-2} (\underset{\sim}{I}-\underset{\sim}{p}(\underset{\sim}{B})\underset{\sim}{1}^T) \; \underset{\sim}{V}[\hat{\underset{\sim}{N}}](\underset{\sim}{I}-\underset{\sim}{1}\; \underset{\sim}{p}(\underset{\sim}{B})^T),$$

so that

$$\underset{\sim}{V}[\underset{\sim}{B}] \doteq (\underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B}) \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{V}[\hat{\underset{\sim}{\pi}}] \underset{\sim}{A}(\underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B}) \underset{\sim}{A})^{-1} . \qquad (3.11)$$

We also have that the covariance matrix for $\underset{\sim}{p}(\hat{\underset{\sim}{B}})$, the estimated cell probabilities, is given by:

$$\underset{\sim}{V}[\underset{\sim}{p}(\hat{\underset{\sim}{B}})] = \underset{\sim}{H}(\underset{\sim}{B}) \underset{\sim}{A} \underset{\sim}{V}[\hat{\underset{\sim}{B}}] \underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B}).$$

The estimators of $\underset{\sim}{V}[\hat{\underset{\sim}{B}}]$ and $\underset{\sim}{V}[\underset{\sim}{p}(\underset{\sim}{B})]$ are similar expressions, where $\underset{\sim}{N}$ and $\underset{\sim}{B}$ are replaced by $\hat{\underset{\sim}{N}}$ and $\hat{\underset{\sim}{B}}$ respectively. These assume that $\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{N}}]$ is readily available. For some problems where q is relatively large compared to r, it would be more efficient to proceed as follows. Let

$$Y_{ki} = 1 \text{ if k-th unit in i-th category}$$
$$= 0 \text{ otherwise,}$$

for k=1, ..., N; i=1, ..., q. Let $\underset{\sim}{Y}_k^T = (Y_{k1}, \ldots, Y_{kq})$, and

$$\underset{\sim}{W}_k = \underset{\sim}{A}^T[\underset{\sim}{I}-\underset{\sim}{p}(\hat{\underset{\sim}{B}})\; \underset{\sim}{1}^T] \; \underset{\sim}{Y}_k.$$

We then obtain:

$$\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{B}}] = (\hat{\underset{\sim}{N}}^T \underset{\sim}{1})^2 (\underset{\sim}{A}^T \underset{\sim}{H}(\hat{\underset{\sim}{B}}) \underset{\sim}{A})^{-1} \hat{\underset{\sim}{V}}(\hat{\underset{\sim}{W}}) (\underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B}) \underset{\sim}{A})^{-1}.$$

We remark that the methodology described in this section can be readily extended to product-multinomial type models, where we have a log-linear model for $\{N_{ij}\}$, but the margins $\{\sum_j N_{ij}\}$ are known.

## 4. DISCUSSION

The techniques described in the paper have been described for some specific models; see, for example, Fuller (1975) and Freeman and Koch (1976). However, the general results are not explicitly described. Many standard statistical packages may be used for the estimation of the parameters of the models described, but the variances and tests of hypotheses given in these packages will not be valid.

The results of this paper depend on the assumption of asymptotic normality of the estimators. Empirical studies on the validity of these approximations are important.

An alternative methodology to estimating many of the parameters described here is given by Imrey, Koch and Stokes (1981, 1982). Their functional asymptotic regression methodology also falls within the general framework described here, with respect to variance derivation and estimation.

## REFERENCES

[1] Frankel, M.R. (1971), Inference from Survey Samples. University of Michigan, Ann Arbor.

[2] Freeman, D.H. Jr., and Koch, G.G. (1976), "An Asymptotic Covariance Structure for Testing Hypotheses on Raked Contingency Tables from Complex Sample Surveys" , Proc. Amer. Statist. Ass. (Social Statistics Section), Part 1, 330-335.

[3] Fuller, W.A. (1975), "Regression Analysis for Sample Survey", Sankya, Series C, 37, 117-132.

[4] Hajek, J. (1960), "Limiting Distributions in Simple Random Sampling from a Finite Population", Publ. Math. Inst. Hung. Acad. Sci., 5, 361-374.

[5] Imrey, P.B., Koch, G.G., Stokes, M.E. (1981, 1982), "Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part I: Historical and Methodological Overview. Part II: Data analysis", International Statistical Review. To appear.

[6] Kish, L., and Frankel, M.R. (1974), "Inference from Complex Samples", J. Roy. Statistic. Soc. B, 36, 1-22.

[7] Madow, W.G. (1948), "On the Limiting Distribution of Estimates Based on Samples from Finite Universes", Ann. Math. Statist., 19, 535-545.

[8] Särndal, C.E. (1978), "Design-Based and Model-Based Inference in Survey Sampling", Scand. J. Statist., 5, 27-52.

[9] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proc. Amer. Statist. Assoc. (Social Statistics Section), 11-18.

[10] Woodruff, R.S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate", J. Amer. Statist. Assoc. 66, 411-414.

# AN OVERVIEW OF CANADIAN HEALTH STATISTICS: PAST, PRESENT AND FUTURE[1]

Lorne Rowebottom[2]

The author briefly reviews the factors determining the production of health statistics in Canada, with particular attention to the different sources of data and to the long-standing co-operation among the many agencies involved in the gathering of health-related information.

Mr. Chairman, I want to express my real pleasure at being a member of this panel because of the opportunity that it affords me to congratulate Dorothy Rice and her colleagues in the National Center for Health Statistics on the occasion of the completion of 25 years of Health Surveys. We in Statistics Canada have long been admirers of NCHS and my congratulations to Dorothy are on behalf of my colleagues in Statistics Canada, particularly those in our Health Division.

Consistent, I hope, with the charge of our Chairman, I have chosen to paint with a very broad brush what seem to me to be trends and determinants of our health which might find echos in other countries and therefore be of interest to this audience.

Two data streams comprise the historic and current sources of Canada Health Statistics. The first is health institutions - predominantly hospitals, both general and mental. From them we derive statistics about a wide range of their characteristics, as well as statistics about their patents and their illnesses. Canadian hospital statistics are amongst the most detailed and comprehensive in the world.

[1] As presented at the American Statistical Association Annual Meeting in Detroit, August 1981

[2] Lorne Rowebottom, Assistant Chief Statistician, Institutions and Agriculture Statistics Branch, Statistics Canada.

The second stream comprises the records generated by registration of births, marriages and deaths from which we derive the critical statistics on causes of death.

A wide variety of statistics is produced from such rich data bases and some important statistics are derived from other sources, for example, those on cancer incidence, from cancer registers, and notifiable diseases. For those who are interested I have a few copies of a Directory of Health Division Information and also I would be glad to send a copy to anyone who wrote to me at Statistics Canada.

The important themes relating to these statistics that I want to touch on this morning are the following:

- First, they measure illness only when individuals seek health care from institutions.
- Secondly, they illustrate the strengths and weaknesses of statistics derived from surveys and from administrative records.
- Thirdly, they represent the availability of information which could only result from a very high degree of co-operation, sustained over a long period of time, between the central agency, federal and provincial departments of health, the institution and hospital associations, and vital statistics registers.

I will return to these three characteristics of the health statistics system: what is measured and what is not, the implications of data sources and the degree of co-operation between the players in the system.

Why have we produced what we have, rather than different products by different means? Looking back over sixty years of health statistics, I found this an interesting question. Assessing how priorities were determined is a judgemental process - just as is deciding on today's priorities. So it is my judgement that in part we responded to changing needs for statistics articulated by users and Royal Commissions, and in part we anticipated changing user needs ourselves and used existing data sources which related to such needs, and because they represented opportunities. They were there to be utilized, like the vein of quartz that a prospector seeks and finds, or stumbles across. In part, we were driven by, and we exploited, the rapidly changing technology. In part the environment of co-operation in which we worked determined what we did. And finally in many parts the resources available to us in terms of dollars, human skills, and data handling capabilities, permitted some things and not others.

These few critical factors:

- articulated and perceived needs,
- data sources available,
- changing technology to process and to analyse data,
- co-operation between players in the system,
- budgets available,

have been the determinants of what we have done. But it will be apparent to you that they are also the determinants of what we are and will be doing.

These forces shift and come together in a changing kaleidoscope so that during one span of time one combination is dominant, to be replaced by another combination.

In Canada all have operated in such ways to bring about significant changes in our health statistics and it seems apparent that there will result even more rapid change. Changing needs should, of course, drive the system and they are in fact doing so, albeit in some respect in an erratic manner. You You will recall my stating that the Canadian measurements of morbidity are largely limited to hospitalized illnesses. This has been widely recognized as a quite unacceptable state of affairs and a few years ago this dissatisfaction led to a federal decision to institute a continuing health status survey of the Canadian population. A survey was carefully planned and tested from both conceptual and methodological points of view. However, only 10 months' data were collected before government-wide budget reductions forced cancellation of the survey. The first results from the data collected have just been published and the data base has shown signs of being a rich research source with significant decision-making implications. Of course, it suffers from the severe limiations of relating to only one point in time. It is too early to state how long it may be before a decision to reinstitute some form of the Canada Health Survey is made. However, I am optimistic that the capacity of such measurements of health status - to throw light on the effects of our lifestyles on our good health and illness, and lead to individual and collective decisions which will affect them - will not be ignored for long.

Let me turn from the area of health-related household surveys where the Canadian track record of responding to changing needs is poor, to one where we have both anticipated and responded effectively to new demands. I refer to epidemiological studies designed to enlighten the kinds of health risks resulting from exposure to various demographic, social, occupational and environmental influences. Thanks to the foresight and persistence of members of our Vital Statistics Staff working with a few other key persons both within and outside Statistics Canada, we have a computer-searchable Mortality Data Base file which includes all deaths in Canada, coded by cause of death, extending back over three decades. We also have a generalized record linkage facility which is being used to link specific exposed population groups to the mortality file. Linkages are also possible to an as yet incomplete but significant ten-year cancer incidence file.

A paper which includes a largely Canadian bibliography on this area will be given by Martha Smith, Head of Occupational and Environmental Health Research Unit, in Scotland before the end of this month. It will be available on request. (Both Martha and John Silins, Chief of our Vital Statistics and Disease Registries Section are in the audience.)

As to other data available to shape the future of Canadian Health Statistics I will only take time to mention the existence of data bases which are very large, potentially very rich, and largely unused for national statistical purposes.

They comprise the administrative records of our national medicare system which record annually in excess of 30 million incidents of primary medical care extended by physicians. We have demonstrated some of the statistical potential of these files and we are now shaping new proposals to develop their use during the next several years. Budgets are expected to be the limiting factor.

New needs should drive the system - new technology does. The influence of computers on health statistics is all-pervasive and is operating to change the availability and uses of health statistics in profound ways.

I want to comment on the use of data - in the form of statistical information, which computers have made possible - by managers, medical personnel and administrators in hospitals, local hospital districts, states, provinces, universities and associations. At federal levels, computers have changed the ways in which data are processed and statistics are used. But in many locations throughout the health community, computers have meant that data are now used for purposes of understanding, for research and for decisions, whereas in the precomputer era they were used little or not at all.

Allowing for some exaggeration - but probably not very much - it was not that long ago when national statistical agencies had almost a monopoly on large-scale data handling capability. What a contrast between then and now when large, fast, sophisticated and easily used information processing capacity is economically available to both large and small organizations. The implications are far-reaching and I suspect not yet fully perceived, but they include at least:

- The existence of many rather than few producers of statistics (many of these will perceive themselves as operators of MIS but statistics is - and will be - the game if not the name.)
- These same organizations will also be much more intensive users of statistics - particularly statistics about their own organizations or jurisdictions.

- As a result there will be greater knowledge of one's own environment.
- There will be greater independence on the part of such organizations and their need - maybe much less perceived need - to rely on others for statistics.
- This ability to utilize the information contained in the administrative records of one's own organization or jurisdiction will almost certainly reduce the tolerance for completing statistical questionnaires, with a resulting increase in the necessity to rely on administrative records. This could result in less information being available about the total environment because of the problems of data comparability between organizations and jurisdictions.

I find it difficult to forecast the impact that these changes will have on co-operation between the many players essential to development and maintenance of a comprehensive and inevitably complex system of health statistics. All I can say is that in Canada - notwithstanding substantial pressures which test and strain the system - co-operation has not diminished. In fact, the reverse is the case and on this score also I am an optimist. I think that one determinant of such co-operation is for national statistical agencies to recognize that their role must change in response to the kind of changes I have described. It is apparent to me that priorities must shift from statistical production to statistical co-ordination.

One final word about what I consider to be an overriding priority, namely, doing statistical analysis of our data bases to determine the messages that are in them, to determine their meaning and significance, and to relate them to the issues and problems confronting us.

For too long, we, at least we in Statistics Canada, have published numbers - myriads of numbers - and failed to translate them into significant indicators. We have left it to others to find the gold in the ore we have mined. I think that we and the health community have paid a high price for our failures (there have been successes) to find the gold, and even shape it into jewellery with which users would enlighten our world, not unlike the way necklaces lend radiance to those who wear them.

# MODELS FOR ESTIMATION OF SAMPLING ERRORS[1]

P.D. Ghangurde[2]

This paper presents results of an empirical study on fitting log-linear models to data on estimates of characteristics and their coefficients of variation (CV) from the Canadian Labour Force Survey. The characteristics were classified into groups on the basis of design effects and models were fitted to data on estimates of characteristic totals and their CVs over twelve month period. The models can be used in situations where estimates of CV are needed for new characteristics, and for providing more precise estimates of reliability of estimates based on past data. The problem of evaluation of fit of the models is considered.

## 1. INTRODUCTION

This paper presents results of an evaluation study on models for estimation of coefficient of variation (CV) of estimates of characteristics based on the Canadian Labour Force Survey (LFS). The LFS is a monthly household survey with a stratified multi-stage area sample design with a sample size of approximately 55,000 households.

Each month estimates of CV are calculated for a set of characteristics using Keyfitz method of variance estimation based on Taylor series approximation [4], [5]. However, computation of appropriate variance estimates for all estimates tabulated from a large scale survey such as the LFS is not possible due to operational constraints of time and

[1] Presented at the American Statistical Association Annual Meeting in Detroit, August 1981.

[2] P. D. Ghangurde, Census and Household Survey Methods Division, Statistics Canada.

costs. The model-based estimates of CV can be used to obtain preliminary estimates of reliability for new characteristics based on the past data, and when estimates of CV for an extended period (e.g. one year) are needed. The models can also be used for obtaining concise estimates of reliability, e.g. alphabetic indicators for ranges of CV.

In section 2 the linear and non-linear models used for estimation of totals and proportions are explained. Sections 3 and 4 review considerations made in forming groups, fitting models and evaluation of goodness of fits.

## 2. THE MODELS

The LFS is a monthly household survey in which dwelling is the final stage sampling unit. Each of the ten provinces in Canada are divided into economic regions which consist of groups of counties with similar economic structure. The economic regions are divided into geographic strata and multi-stage area samples are drawn without replacement with two stages in self-representing strata in the large urban centres and three or four stages in the non-self-representing strata in rural areas. The sample selection in the initial stages is with probability proportional to population size and that in the last stage, in which dwellings are selected from clusters, being systematic.

The design-based estimates within strata are obtained by weighting the data by inverse of probabilities of selection. An adjustment of the basic weight for non-response and ratio estimation within age-sex groups, which are post-strata, is used to obtain final estimates. The census-based population projections for age-sex groups within each province are used as auxiliary variable totals for ratio estimation. More details on the sample design and estimation are given in [5].

The variance estimates of various characteristics at the province level are obtained by Taylor series approximation assuming that the primary sampling units (psus) within non-self-representing strata are selected independently. In self-representing strata the sampled clusters are divided into two groups, which are treated as pseudo-psus and are assumed to have been selected independently. The variance estimate for an estimated characteristic total at Canada level is the sum of corresponding provincial variance estimates [5]. The variance of an estimate $\hat{X}$ of a characteristic total X in a province can also be expressed as

$$V(\hat{X}) = F\ (W-1)\ X\ (1 - \frac{X}{P}), \qquad (1)$$

where P = population for the province,

W = inverse sampling ratio,

F = design effect for the characteristic, and

n = sample size (persons).

The expression (1) for $V(\hat{X})$ relates the variance obtained for the complex ratio estimate based on a stratified multi-stage sample design to the variance of the estimate based on a simple random sample of the same size drawn from the finite population of size P. The sampling variance of an estimate of total based on a simple random sample of size n $(= \frac{P}{W})$ is the usual binomial variance with finite population correction. The term, F, the design effect, represents a factor by which variance is increased due to the effect of such factors as sampling procedure at each stage, the extent of stratification and post-stratification, size of units at various stages and clustering of counts of the characteristic in the province. It may be noted that stratification and post-stratification usually reduce the variance and clustering increases variance of an estimate.

In general, design effects tend to be greater than one due to clustered sample design of the LFS. The labour force status categories such as "employed", "unemployed" by age-sex groups tend to have lower design effects due to post-stratification by age-sex which decreases their variance. Those for labour force status by particular industry tend to be large due to their location in specific areas. Design effects are known to be related to measures of homogeneity and average size of clusters. Models expressing their relationships have been developed for many surveys. In a study on components of variance in the LFS the design effects and measures of homoegeneity have been analyzed for a number of characteristics [2].

A measure of precision of estimates which is independent of the level of the estimate and the scale is coefficient of variation. The $CV(\hat{X})$ is given by

$$CV(\hat{X}) = \sqrt{F(W-1)\ (\frac{1}{X} - \frac{1}{P})}\ . \qquad (2)$$

By taking logarithms to base e on both sides of (2) we have an equation relating CV, X and P given by

$$\log CV(\hat{X}) = \frac{1}{2}\ \log F(W-1) - \frac{1}{2} \log X + \frac{1}{2} \log\ (1 - \frac{X}{P}). \qquad (3)$$

Because of the third term on the right, the equation (3) is not linear in log CV and log X, even if F(W-1) is assumed constant. However, for small values of X the contribution of the third term is negligible. A model based on (3) is given by

$$\log CV(\hat{X}) = A + B \log X + \varepsilon, \qquad (4)$$

where A and B are parameters of the model and $\varepsilon$ is the error term. The estimate of parameter B will differ from $-\frac{1}{2}$ depending on the extent to which B log X approximates $\frac{1}{2} \log [X/(1 - \frac{X}{P})]$ over the range of X. In an evaluation of fits of (4) and of an alternative model (5) given by

$$\log CV(\hat{X}) = A + B \log \frac{X}{(1 - \frac{X}{P})} + \varepsilon, \qquad (5)$$

the goodness of fit for the two models as shown by $R^2$, the ratio of regression sum of squares to total sum of squares, was found to be quite close. The model (4) is linear in log X and log CV and is simpler than model (5).

A non-linear model corresponding to (4) is given by:

$$CV(\hat{X}) = A' X^{B'} + \varepsilon, \qquad (6)$$

where A' and B' are parameters of the model and ε is the error term. The two models (4) and (6) were fitted to data on monthly estimates and their CVs for 90 characteristics in each of 10 provinces and Canada.

## 3. GROUPING OF CHARACTERISTICS

The monthly design effects of LFS estimates for January-December 1980 for each of 90 characteristics excluding total population for each province and Canada were averaged and plotted to decide the ranges for the two groups. In each province, the first group consists of characteristics with design effects greater than D.

Table 1 shows the boundary values D for group I and II in each province and at Canada level, and the number of characteristics in group II. The grouping of characteristics was done by arranging characteristics in increasing order of average design effects. The boundary value D was selected so that the assumption of equal design effects was satisfied as far as possible in group I. The second group consists of all remaining characteristics where the assumption of equal design effects is more crude. Most characteristics pertaining to labour force status by age-sex groups fall in group I. "Employed by industry" and "duration of unemployment" mostly fall in group II. The average design effects differ substantially between provinces and for Canada. More refined grouping of characteristics on the basis of models for design effects is being investigated.

It may be noted that about 80% of the characteristics in each province and for Canada, have been classified in group I. For obtaining a conservative estimate of CV for a new characteristic models based on group II can be used. For a characteristic for which monthly estimates of CV are routinely produced the models for the group in which the characteristic falls, can be used to obtain approximate estimate of CV with a greater precision than that based on monthly data.

In the following section the assumptions made in fitting the models (4) and (6) are explained and model fits are evaluated.

## 4. EVALUATION OF MODELS

The basis of fitting the log-linear model (4) is to treat the model as a simple linear regression model in $y = \log CV(\hat{X})$ and $x = \log X$ and to obtain estimates of parameters A and B in the linear regression framework. The usual assumptions of independence of errors and constant variance have been made. Under these assumptions, $R^2$ provides a measure of fit of the model. The values of the estimated parameters and coefficients of determination, $R^2$, for group I and II in 10 provinces and Canada are given in Table 2. The actual fitting of these models was done by using SAS utility.

All $R^2$ values are significant and quite high indicating that the fits are very good. The error plots do not show any patterns to conclude that the assumption of constant variance is not satisified. Under these assumptions and normality of errors $CV(\hat{X})$ has a log-normal distribution with constant CV for any value of X.

The non-linear model (6) was fitted by Gauss-Newton method using SAS utility. The initial values of parameters $A'$ and $B'$ were assumed to be 1.00 and -0.50 respectively. The number of iterations required to reach convergence was at most 8 for each province and Canada, the convergence criterion being that the relative difference between successive error sum of squares is less than $10^{-8}$. Table 3 shows values of estimated parameters and errors sum of squares for Canada Group II. The errors are approximately normally distributed as shown by normal probability plots.

Since it is of interest to compare the fits of the non-linear model for provinces, Canada and the two groups it is necessary to have a criterion of goodness of fit. In the non-linear model, the total sum of squares is not equal to the total of regression and error sums of squares. A criterion $R'^2$ can be defined as

$$R'^2 = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2},$$

where $\hat{Y}_i$'s are estimated CVs based on the model, $Y_i$'s are observed CVs and $\bar{Y}$ their mean. The summation extends over N, the number of characteristics in the group multiplied by 12, the number of months. In the linear case $R^2 = R'^2$. However, in the non-linear case $R^2 \neq R'^2$ since the total sum of squares is not equal to regression sum of squares plus error sum of squares due to product term not being zero.

The errors $(Y_i - \hat{Y}_i)$ will be small when the fit is good giving a value of $R'^2$ close to 1, the errors $(Y_i - \hat{Y}_i)$ will be large when the fit is poor giving a small value of $R'^2$. When all the points lie on the fitted curve i.e. $Y_i = \hat{Y}_i$ for all i, $R'^2 = 1$. However, in general no lower bound to $R'^2$ seems to exist. The values of $R'^2$ shown in Table 4 tend to be greater for group I as compared to group II, which has 13 to 21 characteristics out of the total of 90.

Although the log-linear model (4) was fitted to data on logarithms of estimates and their CVs and its fit seems to be good, the fitted models for provinces and Canada are used for estimation of CV of estimates. In order to compare the fit of the transformed model to original data of estimates and their CVs, these data and the transformed model corresponding to (4) were plotted for the two groups in 10 provinces and Canada. From these charts it can be concluded that the transformed model corresponding to (4) fits the data of estimates and their CVs better than the non-linear model (6), especially for small values of estimates. The plots of these models for Canada group II are shown on Chart 1 and 2.

## 5. CONCLUDING REMARKS

The characteristics considered are total persons with labour force status by age-sex, industry, marital status and total persons with various ranges of duration of unemployment. However, the models can also be used for proportions instead of totals. The models are not applicable to estimates for subprovincial areas such as urban centres or groups of economic regions, since design effects for these areas are more unstable and can be much higher due to the effect of ratio-adjustment based on projected population at province level [1].

An assumption made in the use of models for a new characteristic is that its design effect is close to the average for the group. This requires finer grouping of characteristics of various types possibly on the basis of models relating design effects with measures of homogeneity for these characteristics. In fitting the models, it was assumed that errors are uncorrelated and that independent variable is fixed. Since twelve monthly estimates for each characteristic were used, there could be correlation in errors for estimates for a given characteristic. Extension of the study to models with errors in independent variable and correlated errors is being considered.

A problem in evaluation of fit of non-linear models, whether actually fitted to data or transformed from linear models, is the lack of a criterion for comparison of fits of different models. The criterion suggested in section 4 may be appropriate for comparison of fits of a model to different data sets, but may not work for different models.

TABLE 1: DESIGN EFFECT BOUNDARY VALUES AND NUMBERS OF CHARACTERISTICS IN GROUPS I AND II*

| Province | Boundary Value (D) | Number of Characteristics | |
|---|---|---|---|
| | | Group I | Group II |
| Newfoundland | 2.3 | 75 | 15 |
| P.E.I. | 1.9 | 73 | 17 |
| Nova Scotia | 1.9 | 74 | 16 |
| New Brusnwick | 2.2 | 77 | 13 |
| Quebec | 1.9 | 73 | 17 |
| Ontario | 1.7 | 69 | 21 |
| Manitoba | 2.0 | 76 | 14 |
| Saskatchewan | 2.8 | 76 | 14 |
| Alberta | 2.1 | 71 | 19 |
| British Columbia | 2.3 | 73 | 17 |
| Canada | 1.9 | 77 | 13 |

* A characteristic belongs to Group I if its design effect (averaged over the 12-month period from January to December 1980) is less than or equal to the boundary value D. If the average design effect is greater than D, then the characteristics is in Group II.

TABLE 2: REGRESSION COEFFICIENTS AND $R^2$ FOR LOG-LINEAR MODEL

| Province | Group | Regression Coefficient A | Regression Coefficient B | $R^2$ |
|---|---|---|---|---|
| Newfoundland | I | 3.3119 | -0.5723 | 0.9534 |
| | II | 3.7757 | -0.6101 | 0.9377 |
| P.E.I. | I | 2.7962 | -0.5617 | 0.9485 |
| | II | 3.1796 | -0.5885 | 0.8887 |
| Nova Scotia | I | 3.4612 | -0.5837 | 0.9702 |
| | II | 3.6412 | 0.5257 | 0.8717 |
| New Brunswick | I | 3.2782 | -0.5545 | 0.9606 |
| | II | 3.7544 | -0.6017 | 0.9357 |
| Quebec | I | 4.3298 | -0.5942 | 0.9686 |
| | II | 4.3093 | -0.5216 | 0.9127 |
| Ontario | I | 4.3825 | -0.6053 | 0.9736 |
| | II | 4.1796 | -0.5009 | 0.9633 |
| Manitoba | I | 3.5155 | -0.5926 | 0.9619 |
| | II | 3.8769 | -0.5640 | 0.9166 |
| Saskatchewan | I | 3.3796 | -0.5700 | 0.9544 |
| | II | 3.5478 | -0.4423 | 0.8994 |
| Alberta | I | 3.6960 | -0.5968 | 0.9678 |
| | II | 3.7526 | -0.5090 | 0.9513 |
| B.C. | I | 3.9847 | -0.5750 | 0.9621 |
| | II | 3.9814 | -0.4708 | 0.8410 |
| Canada | I | 4.3458 | -0.5936 | 0.9703 |
| | II | 4.2357 | -0.5191 | 0.9699 |

TABLE 3: NON-LINEAR LEAST SQUARES: GAUSS-NEWTON METHOD

CANADA (GROUP II)

| Iteration | A' | B' | Residual S.S. |
|---|---|---|---|
| 0 | 1.00000000 | -0.50000000 | 3401.93232121 |
| 1 | 15.22076853 | -0.23647629 | 461.76322678 |
| 2 | 26.47981387 | -0.36743343 | 322.67707190 |
| 3 | 51.94184546 | -0.51147529 | 248.68405130 |
| 4 | 57.29455529 | -0.47434886 | 99.32440727 |
| 5 | 58.32558100 | -0.48419609 | 96.57832290 |
| 6 | 58.28627964 | -0.48409502 | 96.57810754 |
| 7 | 58.28746710 | -0.48409960 | 96.57810746 |

TABLE 4: $R'^2$ FOR GROUP I AND II

| Province | Group | N* | $R'^2 = 1 - \frac{\text{Error S.S.}}{\text{Total S.S.}}$ |
|---|---|---|---|
| Newfoundland | I | 866 | 0.9362 |
| | II | 190 | 0.8835 |
| P.E.I | I | 827 | 0.8925 |
| | II | 294 | 0.7285 |
| Nova Scotia | I | 872 | 0.9790 |
| | II | 192 | 0.7813 |
| New Brunswick | I | 908 | 0.9990 |
| | II | 156 | 0.8639 |
| Quebec | I | 859 | 0.9800 |
| | II | 204 | 0.7804 |
| Ontario | I | 823 | 0.9632 |
| | II | 252 | 0.9208 |
| Manitoba | I | 895 | 0.9691 |
| | II | 168 | 0.8137 |
| Saskatchewan | I | 896 | 0.9436 |
| | II | 168 | 0.8196 |
| Alberta | I | 845 | 0.9701 |
| | II | 228 | 0.8852 |
| B.C. | I | 868 | 0.9319 |
| | II | 204 | 0.7786 |
| Canada | I | 923 | 0.9665 |
| | II | 156 | 0.9286 |

* N for group I can be less than 12 (no. of characteristics) due to exclusion of characteristics with zero estimates.

CHART I

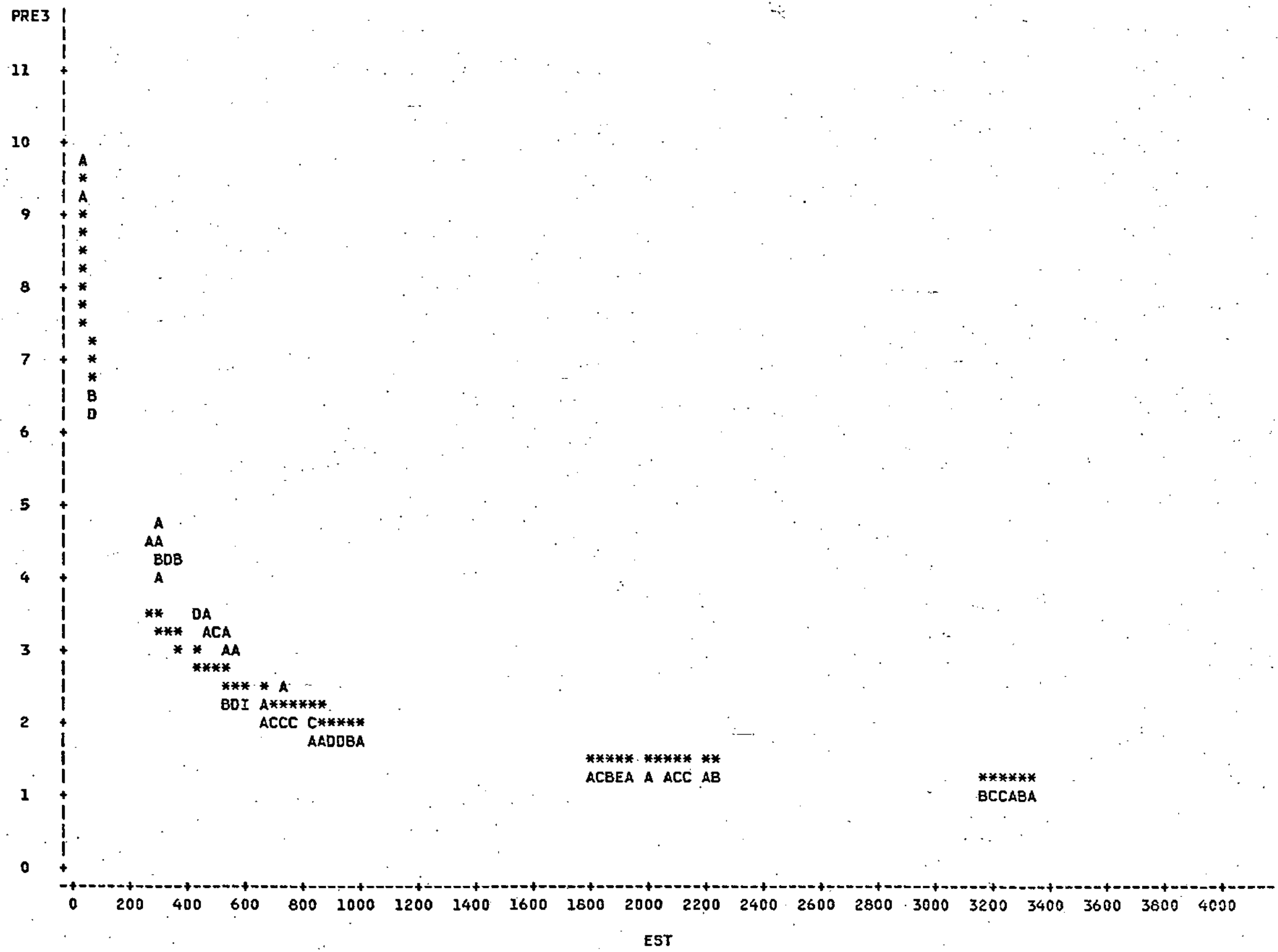
#1 LOG-LINEAR MODEL (transformed)

```
                                   PROV=CANADA   GROUP=2

                      PLOT OF CVHAT1*EST    SYMBOL USED IS *
                      PLOT OF CV*EST        LEGEND: A = 1 OBS, B = 2 OBS, ETC.

CVHAT1 |
       |
       |
11     +
       |
       |
       | *
10     +
       | *
       | *
       | *
 9     + *
       | *
       | *
       | *
 8     + *
       | A*
       | D*
       |  *
 7     +  *
       |  D
       |  B
       |  D
 6     +
       |
       |
       |
 5     +
       |         A
       |        AA
       |         BOB
 4     +         A
       |
       |        **   DA
       |         **   ACA
 3     +          **    AA
       |             **
       |               ****   A
       |                BD* *** A B
 2     +                    AC********
       |                         AADD**
       |                                                      **     A
       |                                                      A**** ***** **
 1     +                                                                                               ******
       |
       |
       |
 0     +
       -+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+---
        0    200   400   600   800  1000  1200  1400  1600  1800  2000  2200  2400  2600  2800  3000  3200  3400  3600  3800  4000

                                                                    EST

NOTE:    98 OBS HIDDEN
```

CHART 2

#3 SIMPLE POWER MODEL
PROV=CANADA GROUP=2

PLOT OF PRE3*EST SYMBOL USED IS *
PLOT OF CV*EST LEGEND: A = 1 OBS, B = 2 OBS, ETC.



NOTE: 101 OBS HIDDEN

## ACKNOWLEDGEMENTS

The author would like to thank the referee for helpful comments.

## REFERENCES

[1] Ghangurde, P.D. and Gray, G.B. (1981), "Estimation for Small Areas in Household Surveys", Communications in Statistics, Theory and Methods, A 10(22), 2327-38.

[2] Gray, G.B. and Platek, R. (1976), "Analysis of Design Effects and Variance Components in Multistage Surveys", Survey Methodology, Vol. 2, No. 1., 1-30.

[3] Kalton, G. (1977), "Practical Methods for Estimating Survey Sampling Errors", Presented at Meeting of the International Association of Survey Statisticians.

[4] Keyfitz, N. (1957), "Estimates of Sampling Variance Where Two Units are selected from Each Stratum". Journal of the American Statistical Association, 52, 503-510.

[5] Platek, R. and Singh, M.P. (1976), Methodology of the Canadian Labour Force Survey. Catalogue 71-526 occasional.

[6] Sprent, P. (1969) "Models in Regression and Related Topics", Methuen and Co.

The Editorial Board wish to thank Mr. R.E. Drover, Chairman, Publication Board and the staff of Administrative Services Division for their continuing support of the production of this Journal.

Acknowledgement is also due to N. Brien, M. Fluet, P. Foy and G. Kriger for their proofreading and preparation of final content.

Thanks are also due to Mrs. D. Edirisinghe, for her patient typing of the Journal, as well as for the execution of numerous other duties associated with its production. Finally, the Editorial Board wish to thank the following persons who have served as referees during the past year.

D.A. Binder
R.G. Carter
G.H. Choudhry
D.P. Dixon
J.D. Drew
S. Earwaker
M. Fluet
P. Foy
G.B. Gray
M.A. Hidiroglou
G. Kriger
S. Kumar
M.L. Lawes
I. Macredie
M.J. March
A. Satin
K.P. Srinath
L. Swain
P.F. Timmons

SURVEY METHODOLOGY

June 1981 Vol. 7 No. 1

A Journal produced by Methodology Staff, Statistics Canada

# CONTENTS