SURVEY METHODOLOGY 1981, VOL. 7 NO. 2

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 finite population parameters, such as the population regression coefficient:

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

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 parameters of a probability distribution (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:

$$H_o: \quad X\beta = \theta_o, \tag{2.1.1}$$

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 super-population parameters) and θ_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 Downing (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 generalized Wald Statistic, defined as:

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

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 non-linear 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:

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

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:

$$P(s|X_1,X_2) = P(s|X_2). \qquad (3.1.2)$$

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:

$$E(Y|X_1) = X_1\beta_1 \qquad (3.1.3)$$

and the design variable X_2 is correlated with Y (conditional on X_1) then the standard OLS estimator of β_1 is not consistent (see Nathan and Holt (1980) and Holt and Smith (1979), who propose modified weighted and unweighted estimates of β_1, which are consistent). Holt, Smith and Winter (1980) give an example of the application of these estimators.

If the linear model:

$$E(Y_i|x_i) = x_i' \beta \qquad (3.1.4)$$

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

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:

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

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 β 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:

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

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:

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

where the weight matrix, $W_n = \text{diag}\ (\Pi_1^{-1}, \ldots, \Pi_n^{-1})$, is the $n\times n$ 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 β, whenever (3.1.4) does hold. It will not, in general, be an optimal estimator of β under (3.1.5) for unequal probability sampling, but will be so if the conditional model variance of Y_i is proportional to Π_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 β, 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 <u>weighted regression</u> (e.g.BMDP) by using the weights $1/\Pi_i$; or from other programmes (e.g. SPSS) by carrying out <u>unweighted regression</u> on the transformed variables $Y_i/\sqrt{\Pi_i}$ and $x_i/\sqrt{\Pi_i}$, but not on the weighted variables Y_i/Π_i, x_i/Π_i. However, it should be noted that under either alternative the <u>reported variances and covariances of the estimators are incorrect</u> and that <u>the standard significance tests (e.g. F tests) are invalid, and can result in grossly misleading conclusions</u>.

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, Π_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:

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

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

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

which is distributed asymptotically as χ^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:

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

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 χ^2 variables with one degree of freedom each.

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

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

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

A conservative test of (3.2.1) can then be obtained by using the statistic X^2/λ_1 in conjunction with a χ^2_{k-1} distribution. λ_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 χ^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 $[\underline{h}(\hat{\underline{p}})]' = [h_{11}(\hat{\underline{p}}), \ \ldots, \ h_{r-1 \ c-1}(\hat{\underline{p}})]$ 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):

$$X^2_{C1} = (n/\hat{\delta}) \sum_{i=1}^{r} \sum_{j=1}^{c} (\hat{p}_{ij} - \hat{p}_{i+}\,\hat{p}_{+j})^2 / (\hat{p}_{i+}\,\hat{p}_{+j}), \qquad (3.2.9)$$

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

$\hat{v}_{ij}(\underline{h})/n$ 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, δ_{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_{C1} 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.