A MODEL-BASED APPROACH TO THE ANALYSIS OF CONTINGENCY TABLES OF DATA FROM COMPLEX SAMPLES

Thomas J. Tomberlin, Harvard University

1. Introduction

The focus of this paper is a proposed method for analysing contingency tables of data obtained from complex samples, which uses a model based approach. The model is linear-logistic as described by Cox (1970), with an addition of random effects terms to allow for clustered sampling schemes. The proposed method could provide for traditional tests of independence and for smoothed estimates of cell totals based on unsaturated models as described for example in Purcell and Kish (1979).

After an introduction to the problem and the need for a solution, we give a brief review of the recent literature on this issue. We follow with a section on the distinction between model-based and design-based inference in finite population sampling. Finally, we consider the proposed method of analysis and conclude with a discussion of the problems of such an analysis and some tentative solutions.

The problem with analysing data from complex samples is that the standard techniques for handling such data are based on an assumption of independent, identically distributed (iid) observations. For samples from finite populations, this assumption is only valid for simple random samples. Such sample designs are rarely if ever used in social surveys. Instead, they tend to be stratified, clustered and many times unequal probability sampling schemes.

That such violations of the iid assumption can lead to erroneous inference in the case of categorical data analysis has been demonstrated using Monte Carlo simulation techniques by Cowan and Binder (1978) and analytically by Fellegi (1978) and Rao and Scott (1979).

When faced with this dilemma, many data analysts have resorted to using the traditional tools based on iid assumptions, and noting that the violation of these assumptions forces one to be extremely cautious about the validity of any consIusions drawn from their analyses. See for example Little (1978).

If a (design-based) variance-covariance matrix for the cell estimates is available, then one solution to some problems of analysis is to use Wald statistics for testing hypotheses based on linear combinations of cell probabilities or logs of cell probabilities. This technique can be used for goodness-of-fit test, tests of independence and tests of parameters of logistic and log-linear models in general. A description of the technique can be found in Grizzle, Starmer and Koch (1968). Examples of its application to the analysis of data from complex samples can be found in Koch, Freeman and Freeman (1975), Freeman and Koch (1976), Freeman, Freeman, Brock and Koch (1976), Freeman, Freeman and Brock (1977) and Tomberlin (1979).

Certainly, the calculation of the variance-covariance matrix of estimates of cell proportion is to be encouraged, but unfortunately, this is not common practice. Indeed, for contigency tables of moderately large dimensions, the size of the associated variance-covariance matrix is so large that the routine calculation and reporting of such matrices seems quite unlikely.

Using models for cluster sampling, Rao and Scott (1979) give approximate methods for testing goodness-of-fit hypotheses which are based on design effects for cell proportion estimates. They argue that information about design effects is more commonly available, even when information about covariances is not available.

Evidently, an approximate solution to the problem of testing hypotheses about independence of classification variables is not so simple. However, after an extensive empirical analysis based on data from the 1971 British General Household Survey, Scott (1980) tentatively reports that unlike the case of goodness-of-fit tests, tests of independence do not seem to be sensitive to the sample design. This is comforting news, but no reason for complacency. The General Household Survey is but one example, and one should not be too quick to generalize from it.

The methods so far discussed pertain only to the problem of testing hypotheses. As we have stated earlier, many times one is interested in estimates of finite population and super-population parameters. Also, these methods are, for the most part, design-based. Rao and Scott (1979) make use of models for the purpose of approximating a function of the eigenvalues of the variance-covariance matrix of the cell estimates. They do not make use of the structure of the population for estimation purposes.

Purcell and Kish (1979) described a method for estimation for small areas which is based on fitting log-linear models to the data and producing maximum likelihood estimates of cell proportions. In a similar vein, Dempster and Tomberlin (1980) proposed a method for fitting logistic models with random effects terms to data from a complex survey for purposes of estimation of census undercount for small areas. It is this method which we describe here.

Before doing so, let us briefly consider the distinction between design-based and model-based inference in the context of finite population sampling.

2. Design and Model-Based Inference

The foundations of inference in the context of sampling from finite populations are presently in a state of controversy. For inference about finite population parameters, such as means and totals, the elements of the controversy, are excellently described in the review paper by TMF Smith (1976) and the discussion which follows it.

In the extreme, the classical sampling statistician would argue that all inference from finite population samples derives from the randomization hypothesis, and thus depends on the sample design rather than on the structure of the population. This notion of inference in finite population sampling dates back to the influential paper by Neyman (1934) and is the basis of much of the development found in the traditional sampling textbooks, such as Cochran (1977) Kish (1964) and Hanson, Hurwitz and Madow (1953). Some authors, such as Kish and Frankel (1974) would argue that the design-induced randomization is of prime (if not sole) importance when estimating regression coefficients and other parameters which most statisticians would regard as model parameters. For example, in their study of different variance estimators, Kish and Frankel (1974) define the regression parameters to be estimated as the least squares solutions which would be obtained from the finite population as a whole if it were available. The sampling variance of estimators of these parameters is defined to be the variance over repeated samples from the finite population. This contrasts with the classical definition of the variance of estimators of regression coefficients which is usually conditional on the observed values of the "carriers" or "independent" variables and is taken over repeated realizations of the model.

On the other hand, there are those who seek to bring finite population inference into the mainstream of statistics by introducing models which attempt to describe the structure of the population. These range from the super-population regression models of Royall (1970), which lead to predictive estimates, to the Bayesian models of Ericson (1969) and Scott and Smith (1969). In the extreme, these approaches can lead to a total rejection of randomization and probability sampling. Indeed, Royall deomonstrates that in the case of estimating the population total for a variable Y, if a super-population model specifying a linear regression of Y on some other variable x, which is known for all elements in the population, is assumed, a purposive selection of the elements having x values at the two extremes of the population will yield an estimator with a smaller variance than could be expected from taking a simple or proportionately stratified random sample. Later he argues that such a strategy is not in general advisable because of possible inadequacies in the model.

By assuming that the sample distribution of the variable x is the same as the population distribution, Royall and Herson (1973) show that the estimate for the population total remains (model) unbiased, even when the model is false. This property is referred to as balanced sampling. Later, Holt (1975) argues that randomization or restricted randomization can lead to samples which are approximately balanced and are thus robust against inadequacies in the super-population model. Ericson (1969), following Savage (1962) before him, argues that randomization can lead to the reasonableness of the assumption of exchangeability which is necessary for his Bayesian inferences about finite population parameters. Neither Holt, in a frequentist super-population framework, not Ericson in Bayesian terms completely justifies the role of randomization in sample surveys. Their arguments appear more designed to justify the use of super-population models or Bayesian inference in spite of the fact that randomization seems necessary.

Rubin (1978) seems to have combined both of these approaches by considering joint Bayesian prior distributions for the finite population variables and the sample itself. Ignoring, for simplicity, the aspects of his paper pertaining to item non-response, this joint prior is given by, $h(X.Y,S) = f(X,Y)\ k(S|X,Y)$. Here, X represents a matrix of data for the finite population which is known for all members of the population. This could be made up of the labels only, or could consist of several auxiliary variables. The matrix Y represents the data, for which values are recorded for sampled units only, and S is a vector indicating the sampled units. S would be a vector of 0's and 1's in the case of a single stage sample, and possibly more complex for a multistage sample.

This factorization of the prior is useful in that $f(X,Y)$ is the prior distribution for the finite population and $k(S|X,Y)$ represents the sampling mechanism. Let x and y be realization of X and S, and $Y = (Y_{(0)}, y_{(1)})$ represent a particular set of observations $y_{(1)}$ for sampled units and the unknown values of Y for unsampled units, $Y_{(0)}$. Rubin shows that if the probability of the observed pattern of sampled units given $(x, y_{(1)})$, $k(S|x,y)$, takes the same known value for all values of the unkown $Y_{(0)}$, then the sampling mechanism can be ignored. In essense, this condition for ignorability means that the population units have exchangeable priors conditional on the $(x, y_{(1)})$. Exchange-ability can be acheived either as the usual Bayesian subjective prior assumption, or as Ericsen would suggest, through randomization.

This ignorability of the sampling mechanism is implicitly assumed on some level by all mode . Thompsen (1978) assumes explicitly that the sampling and model mechanisms are stochastically independent in his discussion of regression analysis from complex samples. T.M.F. Smith (1980) in a paper comparing model-based and design-based inference for regression analysis of complex sample data also touches on the problem of the relationship between the model and sampling distributions.

It would seem that methods which seek to describe both the model and sampling distributions as well as the relationship between the two are likely to be more successful than those which concentrate on one or the other. Completely design-based inference can be very inefficient since it ignores much of the population structure which is often very informative. On the other hand, completely model-based inference is potentially misleading since it ignores the effect of model inadequacies and, as important, ignores possible dependencies between the sampling mechand the model distributions which could distort the distribution of the sample.

Having said all this, for the model-based method of contingency table analysis considered in this paper, we will, like Thompsen (1978), assume that the model and sampling mechanisms are independent. In any practical application, the connection between these two should be explored and included in the model if necessary, possibly using techniques similar to those suggested by Rubin (1978).

Much of the work on finite population inference focuses on the problem of estimating finite population summary measures such as means and totals. When it comes to inference about models and model parameters (e.g. regression, coefficients, independence of classification variables, etc.) the controversy increases. Some, such as Rubin (1978) and Lax (1980) would argue that since models are never entirely accurate, model parameters are never interesting in and of themselves. The finite population (or some expanded definition thereof) is all that should be of interest. Others would argue that model parameters are of interest in themselves both for understanding the relationship between variables and for making predictions in new situations. Econometricians, for example, concentrate on finding "good" estimates of model parameters. See for example Theil (1971). McKennell (1962) and Fields and Tomberlin (1978) describe studies which attempt to build models for the reaction of residents to airplane and railway noise on the bases of social surveys. The purpose of such models is to aid in the siting and construction of possible new airports and rail lines. It would seem that in these cases, model parameters are of interest.

In the case of discrete data analysis, Bishop, Fienberg and Holland (1974) present many instances where hypotheses can be framed in terms of log-linear model parameters. So again, they consider model parameters to be of some interest.

Even among data analysts who agree that parameters are of interest, there is still disagreement. Kish and Frankel (1974) investigate the problem of estimating regression parameters and multiple correlation coefficients from survey data. They define the parameters of interest to be the usual least squares estimates which would be obtained from the entire finite population if it were available. In this respect they argue that such finite population parameters can be regarded as descriptive statistics.

Even some model based data analysts such as Fuller (1975) consider design-unbiased estimates of complete finite population, least squares (model unbiased) estimates of super-population model parameters. In an empirical study by Smith (1980), such procedures are shown to be quite ineffecient when the sampling mechanism can be considered ignorable in the terminology of Rubin (1978). On the other hand, he allows that they are robust against those cases where the sampling mechanism is not ignorable.

In this paper, we consider the parameters of interest to be the (super-population) model parameters and since we assume that the sampling mechanism is ignorable, conditional on other recorded information, we use model-based, maximum likelihood techniques for inference.

3. Random Effects Logistic Models

In this section a method for analysing frequency data from complex samples is proposed which utilizes a logistic model with random parameters. For some time now, logistic models have been used for the analysis of data when the response variable has two categories. The traditional usage of the model is well described by Cox (1970), among others. It is a special case of the log-linear model described by Bishop, Fienberg and Holland (1974), and the techniques described here could also be applied to the fitting of log-linear models.

Dempster and Tomberlin (1980) considered the problem of estimating census undercount for small areas on the basis of a post-enumeration survey (PES) by using logistic models. Since that problem was the motivation for the proposed analytic technique, we will use it here to illustrate the models.

Let the symbol q, with appropriate subscripts, represent the probability that an individual was missed in the census, and let p = 1-q denote the complementary probability of being counted. The subscripts attached to p and q define levels of factors which affect the response. For purposes of illustration we will assume that categories are defined for sex, age groups, and race groups represented by subscripts u, v, and w, respectively. We will represent the triple (u,v,w) by the single symbol μ for convenience. Let us assume we have a three stage survey of individuals in households. Let the symbol ν denote (i,j,k,l), where i represents Primary Sampling Unit (PSU), j represent Secondary Sampling Unit (SSU) within PSU, k represents household within SSU, and l represents an individual within a household.

A typical logistic model might assume the mathematical form

$$\text{logit}(p_{\mu\nu}) = \theta_\mu, \tag{1}$$

where the logit function is defined by,

$$\text{logit}(p) - \ln\{\frac{p}{1-p}\} \tag{2}$$

Note that,

$$\text{logit}(p) = \ln\{\frac{p}{q}\} = -\ln\{\frac{q}{p}\} = -\text{logit}\ (q) \tag{3}$$

The subscript μ on θ_μ in (1) indicates that the logit is allowed to depend on the sex, age, race combination defined by μ=(u,v,w). The absence of any ν-terms indicates that there is no variation in the undercount rates which can be associated with household or areal (ie sampling unit) characteristics once age, race and sex have been introduced to the model. If a model such as (1) without ν-terms is appropriate, then one could analyse the data by the usual techniques for analysis of logistic models.

In the case of measuring census undercount, it seems reasonable to expect that a model like that in (1) would not describe all the variation in undercount rates. It is conceiveable that undercount rates for local areas (SSU's) might vary more than could be explained by differences in sex-age-race compostion alone. One model which would allow for such variation would be

$$\text{logit}(p_{\mu\nu}) = \theta_\mu + \phi j(i) \tag{4}$$

One could speculate on more complicated models than (4), and include the possibility of interaction between the effect of sex, age, and race and geographic characteristics. There are, however, two reasons which make such models difficult, if not impossible to analyse using existing technology.

First, in most cases, particularly the case of national, multi-stage survey such as a PES, the parameter set for models which incorporate effects due to the various stages in the sampling procedure, grows rapidly as the models become more complex. Such large numbers of parameters cannot be handled by classical inference methods, but can be managed by considering them as random In addition, since the sample design is a multi-stage cluster sample, there would be many ν-combinations for which no individuals were observed. Such situations in the context of design and analysis of experiments lead to these parameters being treated as random effects. Not only should this allow for tests of hypotheses regarding these geographical parameters, but, as we shall see, it also leads to empirical Bayes methods of estimation of totals for small subgroups of the population.

The basic idea is to include terms in the logistic model which describe variation in the $p_{\mu\nu}$ within each of the stages of the multi-stage design. Specifically, we may write

$$\text{logit}(p_{\mu\nu}) = \theta_\mu + \phi_{j(i)} + \phi_{k(ij)}, \qquad (5)$$

where the ϕ_i are regarded as drawn from a $N(0,\sigma_1^2)$ population, the $\phi j(i)$ from a $N(0,\sigma_2^2)$ population, and the $\phi_{k(ij)}$ from a $N(0,\sigma_3^2)$ population. These random effects imply that individuals in a PSU have a common element entering into their $p_{\mu\nu}$, and the same occurs for nested classes of individuals in a common SSU and a common household.

Without further research it remains unclear how accurately the variances σ_1^2, σ_2^2, and σ_3^2 can be estimated from sample data, nor is it easy to see, without repeated analyses of the data, what effect different choices of the σ_i^2 would have on final undercount estimates. The models do, however, capture levels of variation which a priori judgement alone strongly suggests must underlie such multi-stage survey data.

Once values of the σ_i^2 are tentatively adopted, it becomes possible to introduce corresponding factors into the likelihood analysis, and hence produce approximate normal posterior dustributions for the logit($p_{\mu\nu}$) which automatically and correctly weight undercount frequencies observed at the various levels of the multistage design. For example, the posterior mean of logit($p_{\mu\nu}$)for an individual l(ijk) who appears in the PES automatically uses information from the individual's household, SSU, and PSU. More remarkably, a posterior mean logit($p_{\mu\nu}$) can be found for an individual l(ijk) not in the PES, and again the PES counts are automatically weighted, where the weighting scheme depends on which if any among i, j(i), or k(ij) appear in the sample. Similarly, we can find posterior variances which appropriately incorporate the available information about each individual.

The basic mathematical development facilitating approximate computation of the required posterior means and variances appears in Laird (1975). Some initial experience with variance estimation is found in Miao (1977). Neither of these papers treats examples of the degree of complexity required for a real multi-stage survey so that detailed research and development will be needed, but the principles are in place.

The problem of testing hypotheses under this framework needs even more work. One could imagine making interval estimates of the variance components σ_i^2 for model (5), and noting whether they included the origin. This kind of procedure whould produce reasonable results for the purposes of model fitting. As is usual for models with random effects, if a variance component is not different from zero, then the corresponding set of random parameters should be dropped from the model. The tools for hypothesis testing require much more research.

4. Conclusion

In this paper, we have presented an outline for a means of analysing categorical data obtained by complex sampling schemes. It should be emphasized that this is a proposal, and much remains to be done before it is established as a workable option for data analysts.

Once the technology is in place, there will remain at least one obstacle to the widespread adoption of the techniques. Data from many surveys, particularly those conducted by government agencies, are widely used by groups other than those who actually design and implement them. Confidentiality problems alone would make it very difficult, if not impossible for the detailed micro-data necessary for this type analysis to be provided to all users. Indeed, most users would have neither the inclination nor the ability to carry out such an analysis. On the other hand, as the technology does become avialable, it behooves these survey organizations to carry out some of these analyses on their own. This would enable them to warn or reassure users as to the effects of clustering in the sample on ordinary contingency table analyses. Hopefully, observations similar to those reported by Scott (1980) regarding data from the General Household Survey in Britian will apply to most other such surveys.

Our observations on the relationship between the sample design and the model should have some impact on sample design. If one of the major purposes of the survey is to provide data for complex analyses, then the design should be planned so as to facilitate these ends. Specifically, data generated by sampling schemes which can be treated as ignorable in the sense intended by Rubin (1978) are much simpler to analyse.

REFERENCES

1. Bishop, Yvonne, Stephen Fienberg and Paul Holland (1974), Discrete Multivariate Analysis, Theory and Practice, Cambridge, MA: MIT Press.

2. Cochran, William (1977), Sampling Techniques, 3rd ed. New York: John Wiley and Sons.

3. Cowen, J. and Dave Binder (1978), "The effect of a two-stage sample design on tests of independence of a 2x2 table", Survey Methodology,4.

4. Cox, D. R. (1970), The Analysis of Binary Data, London: Methuen.

5. Dempster, Aurthur D. and Thomas J. Tomberlin (1980), "The analysis of census undercount from a post-enumeration survey", presented to the Conference of Census Undercount, Arlington, VA.

6. Ericson, W. A. (1969), "Subjective Bayesian models in sampling finite populations", J.R. Statist. Soc. B, 31.

7. Fellegi, Ivan (1978), "Approximate tests of independence and goodness-of-fit based on stratified, multi-stage samples", Survey Methodology. 4.

8. Fields, James M. and Thomas J. Tomberlin (1978), "Noise survey design and the precision of statistical results: further evaluation of the design of a national railway noise survey", Proceedings of Internoise 78, San Francisco.

9. Freeman, D., J. Freeman, and Brock (1977), "Modularisation for the analysis of complex sample survey data", Invited paper, Int. Assoc. of Survey Statists., ISI meetings, New Delhi.

10. Freeman, D. J. Freeman, Brock and G. Koch (1976) "Strategies in the multivariate analysis of data from complex surveys II", Int. Statist. Rev. 44.

11. Freeman, D. and G. Koch (1976), "An asymptotic covariance structure for testing hypotheses on raked contingency tables from complex surveys', Amer, Statist. Assoc., Proceeding of the Social Statistics Section, Boston.

12. Fuller, Wayne A. (1975), "Regression analysis for survey sampling", Sankhya C, 37.

13. Grizzle, J., Starmer and G. Koch (1969), "Analysis of categorical data by linear models", Biometrics.

14. Hansen, M. H., W. N. Hurwitz and Madow (1953) Sample Survey Methods and Theory, New York: John Wiley and Sons.

15. Holt, D. (1975) "A Generalization of Balanced Sampling" Bulletin of the Internation Statistical Institute, Vol. XLVI, Book 3.

16. Kish, Leslie (1964), Survey Sampling, New York: John Wiley and Sons.

17. Kish, Leslie and Martin Frankel (1974), "Inference from Complex Samples", J.R. Statist. Soc. B, 36.

18. Koch, Gary, D. Freeman and J. Freeman (1975), "Strategies in the multivariate analysis of data from complex surveys", Int. Statist. Rev., 43.

19. Laird, Nan (1975), "Log-linear models with random parameters: an empirical Bayes approach", Ph.D. Dissertation, Department of Statistics Harvard University.

20. Lax, David (1980), Private communication.

21. Little, R. J. A. (1978), "Generalized linear models for cross-clissified data from the WFS", World Fertility Survey Technical Bulletin No. 5/ Tech. 834, London.

22. Miao, Lillian (1977), "An empirical Bayes approach to analysis of inter-area variation", Ph.D. Dissertation, Department of Statistics, Harvard University.

23. McKennal, A. C. (1963) "Aircraft noise annoyance around London (Heathrow) airport", Central Office of Information. SS 337, London.

24. Neyman, J. (1934), "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection", J. R. Statist. Soc., 97.

25. Purcell, N. J. and Lelie Kish (1979), "Estimation for small domains", Biometrics 35.

26. Rao, J. N. K. and A Scott (1979), "The analysis of categorical,data from complex sample surveys I: chi-squared tests for goodness-of-fit" Amer. Statist. Assoc., Proceedings of the Section on Survey Research Methods, Washington, D.C.

27. Royall, Richard (1970) "On finite population sampling theory under certain linear regression models", Biometrika 57.

28 Royall, Richard and Herson (1973), "Robust estimation in finite population, I", J. American Statistical Association,68.

29. Rubin, Donald B. (1978), "The phenomenological Bayesian perspective in sample surveys from finite populations: foundations", presented to the Institute of Math. Statist. meetings, Rutgers University.

30. Savage, L. J. (1962), The Foundation of Statistical Inference, London: Methuen.

31. Scott, A. J. (1980), "Chi-squared tests for the analysis of categorical data from complex surveys", presented to the Symposium on Survey Sampling, Carleton University, Ottawa, Canada.

32. Scott, A. J. and T. M. F. Smith (1969), "Estimation in multi-stage surveys", J. of the Amer. Statist. Assoc. 64.

33. Smith, T. M. F. (1976), "The foundations of survey sampling: a review". J. Roy Statist. Soc A 139.

34. Smith, T. M. F. (1980), "Regression analysis of survey data", presented to the Symposium on Survey Sampling, Carleton University, Ottawa, Canada.

35. Theil, H. (1971), Principles of Econometrics, New York: John Wiley and Sons.

36. Thompsen, Ib (1978), "Design and estimation problems when estimating a regression coefficient from survey data", Metrika 25.

37. Tomberlin, Thomas J. (1979), "The analysis of contingency tables of data from complex samples", Amer. Statist. Assoc., Proceedings of Section on Survey Research Methods, Washington, DC