ANALYSIS OF COMPLEX SURVEYS

A thesis presented in partial fulfillment

of the requirements for the degree of

Masterate in Science

in Statistics

at Massey University

JANE YOUNG

May 1997

ACKNOWLEDGEMENTS

I can't believe that it is finally coming to an end!! Of course, the first person I must thank is Associate Professor Stephen Haslett. Thank you Steve for that endless supply of time, effort, supervision, guidance, advice, wisdom and support over the year (oh, and patience!!). I have learnt so much from Steve, maybe because he never wanted to tell me the answers. '...I could tell you the answer but you would not learn anything...' was one of his favourite sayings I seem to recall. His endless supply of knowledge never ceased to amaze me.

Thanks also goes out to Dr Siva Ganesh for the use of his most beloved PC. Without it I fear that my computer analyses would have taken me another year to do (when they finally let me have a bigger and faster machine!). Also thanks for his expertise in SAS and multivariate statistics, which came in handy at almost the right times.

I also must thank Mr Alasdair Noble for the time and effort which he put into reading what I thought was my final draft. Even though it meant more work and sleepless nights, his comments and suggestions were invaluable.

The staff in the Department of Statistics have been a wonderful support throughout my studies at Massey University. Even though I am absolutely sick and tired of studying, I know I will miss this place when I finally leave.

Lastly, I would like to thank all of those friends and family of mine that have supported me and kept asking me 'when are you going to finish?'. I'm not sure how many of them actually knew what I was studying though.

WHEW!! Well, all I can say is that the light at the end of the tunnel is no longer an oncoming train!

ABSTRACT

Complex surveys are surveys which involve a survey design other than simple random sampling. In practice sample surveys require a complex design due to many factors such as cost, time and the nature of the population.

Standard statistical methods such as linear regression, contingency tables and multivariate analyses are based on data which are independently and identically distributed (IID). That is, the data is assumed to have been selected by a simple random sampling design. The assumptions underlying standard statistical methods are generally not met when the data is from a complex design. A measure of the efficiency of a design was found by the ratio of the variance of the actual design over the variance of a simple random sample (of the same sample size). This is known as the design effect (deff). There are two forms of design effects; one proposed by Kish (1965) and another termed the misspecification effect (meff) by Skinner et al. (1989). Throughout the thesis, the design effect referred to is Skinner et al. (1989)'s misspecification effect. Cluster sampling generally yields a deff greater than one and stratified samples yields a deff less than one.

Some researchers have adopted a model based approach for parameter estimation rather than the traditional design based approach. The model based approach is one which each possible respondent has a distribution of possible values, often leading to the equivalent of an infinite background population,

called the superpopulation. Both approaches are discussed throughout the thesis.

Most of the standard computing packages available have been developed for simple random sample data. Specialized packages are needed to analyse complex survey data correctly. PC CARP and SUDAAN are two such packages. Three examples of statistical analyses on complex sample surveys were explored using the specialized statistical packages. The output from these packages were compared to a standard statistical package, The SAS System. It was found that although SAS produced the correct estimates, the standard errors were much smaller than those from SUDAAN. This led, in regression for example, to a much higher number of variables appearing to be significant when they were not.

The examples illustrated the consequences of using a standard statistical package on complex data. Statisticians have long argued the need for appropriate statistics for complex surveys.

CONTENTS

CHAPTER THREE

CHAPTER FOUR

APPENDIX

LIST OF FIGURES AND TABLES

PREFACE

This thesis covers some standard methods for analysing complex surveys. Chapter one discusses some common sampling designs and provides general theoretical background to the problem of analysing complex survey data.

Data from two survey questionnaires involving some complex design are used throughout the thesis to illustrate statistical methods and to provide some actual survey data for analyses. The questionnaires used are presented in chapter two.

Chapter three discusses the effect of a complex design in regression analysis. A brief overview of the traditional regression methods is given and this leads to the effect of a complex design on the regression parameters. To adjust for the survey design, alternatives to the ordinary least squares estimator are considered.

Another common statistical technique in sample surveys is the use of contingency tables for categorical data. Analysis of contingency tables in chapter four includes various chi-square test statistics, the effect of complex designs on the standard chi-square statistics and the development of appropriate adjustments.

In chapter five, the focus is on multivariate data analysis. The effect of complex designs on the covariance matrix and different estimates of the covariance matrix is considered under the design and model based approaches. In particular, principal components is discussed as the main multivariate technique.

Some computing examples based on 'real life' sample surveys are in chapter six. The computing programs used for the analysis of a complex survey are PC CARP and SUDAAN. The outputs from these packages are compared with a package that does not adjust for complex surveys; this package will be SAS.

The final chapter includes a summary and conclusions.