STATISTICAL ISSUES IN LARGE COMPLEX HOUSEHOLD SURVEYS CONDUCTED IN DEVELOPING COUNTRIES: THE LESOTHO NATIONAL HOUSEHOLD HEALTH AND NUTRITION SURVEY

By

Mark Joel Paiker

A Thesis Submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy in the Faculty of Science, University of the Witwatersrand, Johannesburg.

Johannesburg 1998.

ABSTRACT

The statistical issues in conducting a National Household Survey such as the Lesotho National Household Health and Nutrition Survey (NH&NS) are considered from two aspects:

1) Issues related to data processing and the management of complex survey designs such as the NH&NS and 2) issues related to the weighting and analysis of such data.

The problems with short-term consultants, the need for improved management skills and quality control are addressed in this thesis. To prevent common data errors found in a survey such as the NH&NS and to speed up the release of results, recommendations of improved data collection techniques are made. Since developing countries often lack the capacity to conduct large national surveys, the proposal of a well-conducted single round instead of a multi-round survey is investigated. Calculating weighting factors can be a complicated and time- consuming process for a Bureau of Statistics in a developing country, which might prevent the early release of the results. Therefore it is shown how appropriate weighting factors for a complex survey design are generated. It is also necessary to consider the effect of clustering and stratification, whether it is for presenting results as confidence intervals or for fitting models to the data. By applying the technique of portability to surveys such as the NH&NS, much time can be saved and the results released far earlier than when more convential analysis techniques are used. Chi-square analyses of two-way tables, log-linear modelling and logistic regression are the most common analytical techniques applied to the data from such surveys. The adjustments to these techniques to compensate for a complex survey design are demonstrated. These techniques are applied to the child nutrition, maternal care, disability and injury sections of the NH&NS, with an emphasis on the child nutrition data. Various programs have been written in SAS to perform these analyses. The results generated in this thesis using these techniques should be useful to the Lesotho Ministry of Health and serve as a reference for similar health and nutrition surveys in the future.

Declaration

I declare that this thesis is my own unaided work. It is submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any other degree or examination in any other university.

Mark Joel Paiker

14 day of JANUARY, 1998.

To my parents Mike and Sharon Paiker
With sincere thanks for their love, encouragement and for always being there for me.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation for the support and assistance I received from my family, friends and colleagues who made this work possible.

I am particularly indebted to my supervisors Professor Paul Fatti, currently Head of Department of Statistics and Actuarial Science, University of the Witwatersrand, and Professor Max Price, currently Dean of the Health Sciences, University of the Witwatersrand, for their enthusiasm, patient guidance and encouragement in assuring the completion of this work.

I am very thankful for the assistance of the Director and staff at the Bureau of Statistics in Maseru, Lesotho, particularly of Ms. Botsoa Hloaele, for the tireless effort she made in cleaning the data.

I am also very grateful to Professor Stoker for his guidance and for encouraging me to pursue the field of complex survey designs.

Finally, I wish to thank my colleagues Mr Eugenius Senaoana and Mr Andrew Samson for all their encouragement and support in completing this thesis.

TABLE OF CONTENTS

CHAPTER 5

Dealing with Anthropometric data

Page

CHAPTER 6

The estimation and portability of standard errors and design effects of ratio estimates using data from a complex survey design

CHAPTER 7

The analysis of two-way and multiway tables

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 HEALTH INTERVIEW SURVEYS IN DEVELOPING COUNTRIES

Health interview surveys are the most common and preferred method of collecting health information in developing countries. Other sources are through a population census, records from medical examinations, and disease registers. The call for "health-for-all" by the year 2000 (World Health Organization 1981) has had a strong influence on the continuation of the interview survey as the primary source of health information. In order for a country to plan and manage its health services it is necessary to continually update its information on various health factors such as morbidity, nutrition, and maternal care. In addition to this the attitude and knowledge of the population towards health services would also need to be considered when the required preventive and curative measures are made. Even though health interview surveys have been conducted for the past few decades, and much has been written in the literature about the various surveys around the world, each country seems to experience its own unique problems in the management of these surveys. In developing countries where capital·to fund surveys is limited, it is imperative to ensure that they are conducted effectively and efficiently. Besides producing accurate results for that country, it would also encourage the donor country to continue with its financial support.

One could question the wisdom of conducting expensive national health surveys in preference to directly funding health services. The main argument in favour of health surveys is that they are necessary for the planning and management of the health services. As K L White (1985) states: "If the 'received wisdom' and 'best guesses' of those responsible are inadequate for establishing the dimensions of suffering and the numbers and locations of the people experiencing the suffering, then a survey may be helpful for ordering the problems and populations in some rational fashion." Health planners require knowledge about the frequency and trends of morbidity, nutrition, maternal care, etc., as well as the attitudes, accessibility to, and utilization of the health services by the population. Survey findings enable them to monitor the effectiveness of the existing services, motivate the necessity for the establishment of new health services, as well as place more emphasis on the prevention and treatment of specific diseases and illnesses. Therefore in order to justify the large amounts of money spent on these surveys, it is essential that once the survey is approved and carried out, the health planners receive reliable information from a well executed survey within a short period of its completion.

All sample surveys are subject to both sampling errors and non-sampling errors. Improvements in the design and implementation of these sample surveys in recent years have helped to minimize these errors. However health surveys in developing countries are still prone to incomplete or partial reporting, response errors and to a lesser degree coverage errors (Croft (1991)). Developing countries have had to rely predominantly on the experiences of industrialized countries when developing suitable methods for their surveys. However many situations in the survey process are unique to developing countries. Kroeger (1985) suggests reducing the sample size of a national survey in favour of a more intensive small-scale survey since "Supervision requires much travelling, which is extremely uncomfortable and difficult in most rural areas of the developing world. Compromises between the intensity of supervision and the feasibility of travelling have to be found."

Timaeus et al (1988) had the following reservations about surveys in developing countries: "Many of these surveys are unsystematic, poorly planned and give insufficient attention to the development and pretesting of questionnaires and procedures. Many previous surveys have been slow to produce results and, when they have appeared, have not addressed the questions that policy makers need answering. Serious problems are being caused by the unintegrated, conflicting, duplicating and expensive (in time, labour, and money) information-gathering activities already taking place in many countries, especially in Africa. The situation has been aggravated by frequent revision of existing procedures and introduction of new ones, often being imposed on countries through demands for information from the international agencies."

1.2 COMPLEX SAMPLING

If each element of the sampled population has a known positive probability of being selected in the sample, then the sampling procedure is known as a probability sampling procedure. If a non-probability sampling procedure is used, one cannot in general comment on the unbiasedness of the estimators and on their standard errors. The four basic probability sampling procedures are simple random sampling, systematic sampling, stratified sampling, and cluster sampling. These four sampling procedures are usually combined in practice to form a multi-stage stratified cluster sampling design, known as complex sampling.

Simple random sampling is the procedure whereby, say, n elements are randomly drawn from the sampled population in such a way that each element of the population has the same probability of being included in the sample. There are three problems associated with simple random sampling:

(1) An entire list of the sampled population is necessary. This list is not always available, and it may be difficult and expensive to construct.

(2) One cannot guarantee that the sample will be representative of the population being sampled. For example if one were sampling 15 primary schools around South Africa, and all the schools randomly chosen were from the Western Cape area, these schools would not be representative of South Africa.

(3) If the elements selected in the sample are geographically widely spread over an area, it will be expensive and time consuming to collect the data.

A **systematic sampling procedure** is one whereby the elements are drawn in some regular way, starting from a random starting point, from a complete list of the elements of the population. If correctly approached and executed, this procedure has the potential of producing samples that are representative of the population. The danger with systematic sampling is the possible presence of hidden periodicities, where every n^{th} element chosen has a similarity that could bias the results.

Stratification is the process whereby the population is divided into subpopulations or strata. An independent sample is then drawn from each of the strata. In national household surveys the stratification variables are usually the geographic area or region and degree of urbanization (rural or urban). The reasons for stratification are (a) to ensure the representation of each segment of the population in the sample and (b) to reduce the standard error of the estimators of the population parameters by dividing the strata in such a way that each stratum is more homogenous than the population with regards to the relevant variable(s) or characteristic(s) under investigation.

Due to the costs and logistics of randomly selecting households or respondents within a stratum, it is often necessary to sample clusters within the strata. **Clusters** are usually formed by grouping geographically neighbouring elements of the population. The more heterogenous with respect to the variable(s) or characteristic(s) under investigation the households of the population grouped within clusters are, the smaller the standard errors of the population estimates are. Since these elements tend to be alike there is usually a positive intracluster correlation, which results in the increase of the standard errors of the estimators of the population characteristic(s) or variable(s). For a fixed total sample size the smaller/greater the cluster size, the smaller/greater is the standard error of the estimators. A cluster size of one is the same as using a simple random sample. The problem with small cluster sizes is that it results in higher costs. Therefore a compromise is necessary in deciding between lower costs or lower standard errors of the estimators.

Nearly all surveys in developing countries are stratified cluster samples, with clusters comprising groups of households in each of the regions or strata. Since the observations from such a design are not independent and identically distributed (iid), it is necessary to make adjustments in order to analyze such data. The standard statistical packages assume iid observations drawn from infinite populations and hence are not appropriate for most survey designs.

The recommendations and proposals in this thesis are based on my experience of the Lesotho National Household Health and Nutrition Survey (NH&NS) that was conducted over four survey rounds in 1988 and 1989, and generalized wherever possible to any other developing countries conducting similar surveys. The first round data is not considered in this thesis due to the various problems encountered in the data processing phase (see Chapter 3 for more details).

This thesis can be divided in to two main sections:

(i) Issues related to data processing and management of complex survey designs such as the NH&NS.

It is necessary to address the concerns made above by Timaeus et al. (1988) concerning the capabilities of developing countries in conducting health interview surveys. Given the limited capacity of the local staff of developing countries to conduct large scale surveys, the unique survey environment encountered in developing countries (i.e. the extensive rural areas, illiteracy, poor administrative records, and the lack of adequate training in conducting surveys), and the fact that low health service utilisation makes interview surveys an essential source of health information, it would seem that a modified approach to the survey analysis process should be implemented. Some proposals towards this end are made in this thesis.

Various recommendations have been made in the literature to improve the methodological approach of health interview surveys in developing countries. These recommendations are concerned with the sampling design, the questionnaire design, the questionnaire content, as well as the general fieldwork of the health interview survey in developing countries (Kroeger (1983) and Ross and Vaughn(1986)). These are not repeated in this thesis. However, less has been written about the overall management of surveys, the use of consultants, and particularly the mechanisms for improving the quality of data processing (editing, entry, cleaning and analysis).

Based on my experiences and the experiences of the survey manager of the NH&NS, the influence of short-term consultants and the need for improved management skills and quality control are addressed. Common data errors found in a survey such as the NH&NS are indicated. To prevent errors and speed up the release of results, data collection techniques using the latest computer technology, are considered. Based on the lack of capacity in a developing country to conduct large national surveys, the proposal of a well conducted single round instead of multi-round surveys is investigated. A data cleaning procedure used in the NH&NS for child anthropometric data is also described.

(ii) Issues related to weighting and the analysis of such data.

Calculating weighting factors can be a complicated and time consuming process for a Bureau of Statistics in a developing country, which might prevent the early release of the results. Strictly speaking, while it is necessary to compensate for missing households (unit nonresponses), it might not affect the outcome of the results significantly if a weighting factor is not incorporated into the final weight to compensate for missing items (item nonresponse). The weighting procedure employed for a complex survey design such as the NH&NS is firstly described, and thereafter the question of excluding the weighting factor to compensate for missing item responses is addressed.

The effect of using an incorrect standard error, based on a simple random sample instead of a complex survey design, is demonstrated (Lehtonen and Pahkinen 1994). In the case of a longitudinal survey such as the NH&NS the possibility of carrying over information concerning the standard error from one survey round to the next is considered.

Two-way table analysis, log-linear modelling and logistic regression are the most common analytical techniques applied to the data from such surveys, and the adjustments to these techniques to compensate for a complex survey design are demonstrated. These techniques are applied to the child nutrition, maternal care, and morbidity sections of the NH&NS. It is hoped that the results generated in this thesis using these techniques will be found to be useful for the Lesotho Ministry of Health and serve as a reference for similar health and nutrition surveys to be conducted in the future.

Various programs have been written in SAS (SAS(1994)) to facilitate the investigation of some of the abovementioned issues, and to analyze the data where such programs are not readily available. The source code as well as a description of how to run these programs can be found in the appendices.

1.3 CONTENTS OF THE THESIS

Below is a short description of the contents of the thesis which deals amongst others, with some of the issues raised in Sections 1.1 and 1.2.

Chapter 2 gives a description of the NH&NS. This incorporates the sampling design, the type of data collected, the original objectives of the survey, and my role in the survey.

Due to numerous data errors and missing information that appeared in the NH&NS, it was necessary to implement a data cleaning procedure to correct data errors that occurred in the editing or data entry phases. In Chapter 3 these errors will be identified and explained. Recommendations will be made to prevent similar errors from reappearing in future surveys. Recommendations for a computerized data entry process are also made. It has been suggested that single round surveys on a smaller scale should be used in developing countries until the necessary expertise is attained (Kroeger 1983). In Chapter 3 a comparison of the results is made between the entire data set and the three rounds separately, in order to observe whether similar results are observed in spite of the reduced sample size. Time constraints and the fact that the fieldwork phase had been completed, made it impossible to have callbacks. Therefore not all the data could be cleaned. The cleaning was restricted to the variables analysed in subsequent chapters.

Weighting is required in most national surveys for three reasons. Firstly population totals of the various variables are required from the survey, and therefore weighting to the most recent Census is required. Secondly stratified samples must be adjusted where sampling probabilities in different strata are not equal, and thirdly weighting is necessary to compensate for missing information, so that biases are reduced from the estimation and analysis phases. As in most national health surveys there is located, or respondents not knowing the answers to the questions. Data might be missing for a single question, an individual, an entire household, and in some rare cases even an entire village (cluster). The most recent Census figures are used to determine the weighting factors described above. This weight is multiplied by a factor which takes into consideration the "raising factor" which adjusts the data according to the age/sex distribution of the Census, as well as the population growth rate since the last Census. The final weight also incorporates the weighting factor for the compensation of item nonresponses. Imputation can also be used to compensate for item nonresponses, so the various imputation techniques are considered, and an example applying this technique to the NH&NS is demonstrated. In Chapter 4 SAS (SAS (1994)) programs which were written for the imputation and subgroup weighting procedures of this thesis, are discussed and referred to in the Appendices. In Chapter 4 the weighting process in the NH&NS is explained, and a resampling comparative study of two weighting factors, one incorporating the factor for compensating for item nonresponses and the other without it, and imputation is made. Applying a resampling procedure, the effect of varying weighting factors on the standard error will be investigated by comparing the standard errors of a data set with missing observations, with and without subgroup weighting and with imputation.

In Chapter 5 the child nutrition section of the survey is considered. Various problems from the literature associated with these anthropometric measurements are discussed. Data cleaning procedures that were designed for the NH&NS to identify systematic erroneous weight and height entries are explained.

Most national health interview surveys make use of a complex sampling design which is made up of, among others, stratification and cluster sampling, as described in Section 1.2. In Chapter 6 methods are discussed for estimating standard errors and design effects of proportions from a survey, taking the effect of complex sampling and weighting into consideration. The design effect is the ratio of the sampling variance of an estimator wher. the complexity of the design is taken into account, to the sampling variance of the estimator if it is assumed that the sample data have been obtained by means of simple random sampling. Since well known packages such as SAS and SPSS (SPSS X (1995)) do not have procedures for estimating standard errors and confidence intervals that take the complexity of the survey design into consideration, a prograr. has been written in SAS to serve this function. The procedure of 'portability' is also considered in this chapter. This procedure uses information from the same or similar surveys to estimate the standard errors, the rate of homogeneity, and the design effect of each cell in a multiway table. Applying this procedure effectively and carefully could save a Bureau of Statistics in a developing country like Lesotho much time and ensure quicker analysis and release of results.

Since most of the data from surveys such as the NH&NS are categorical, the two most appropriate techniques considered in this thesis are log-linear modelling and logistic regression. In Chapter 7 the log-linear model is applied to the child nutrition data of the NH&NS, taking the complexity of the survey design and the weighting into consideration. This analysis entails including a correction factor based on the design effect for each cell in a multiway table, and rescaling the multiway table to compensate for weighting. A question asked in this chapter is whether the exclusion of the correction factor will materially affect the conclusions of the analysis. Another approach using the SUDAAN(1994) statistical software system, where the full covariance matrix of the weighted proportions is used to calculate the Wald statistic which tests the adequacy of the log-linear model, is also described. The interpretations of the results requested by the Bureau of Statistics(BOS) of Lesotho for the child nutrition data, are presented at the end of this chapter. To determine associations (requested by BOS) between malnutrition and various demographic variables, two-way table analyses are presented. A comparison of the 'correction factor' analyses are presented. A comparison of the 'correction factor' technique with the 'full covariance' techniques is also made. In the case where a log-linear model is developed, there is no specific dependent (response) variable.

In Chapter 8 the logistic regression model is considered. The process of developing the best fitting model to determine malnutrition is described in this chapter. Stratification and clustering under the conditions of this model are discussed and applied to the NH&NS data. SAS programs are written to test the goodness-of-fit of the model.

Using the techniques from the previous chapters, the remaining sections of the NH&NS are analyzed in Chapter 9, i.e. disability, injury, and maternal care. These results can provide the Lesotho Ministry of Health with valuable baseline information (where certain tables are generated for the first time), assist in monitoring trends, and they can evaluate the effectiveness of intervention programs instituted by the Ministry of Health.

The last chapter (Chapter 10) consists of a summary and conclusions based on the contents of the previous chapters.

1.4 COMMENT

At present most developing countries have a central bureau of statistics which have the infrastructure to undertake such surveys. These bureaux have staff in the various fields of the survey process i.e. interviewers, editors, data processors and supervisors, as well as the facilities to process, print and publish the results. Many developing countries employ overseas consultants to manage their surveys. These consultants are usually employed on a temporary basis and their stay usually lacks continuity. It is necessary therefore that the local bureaux of statistics eventually become capable of conducting their surveys independently, so that they do not become reliant on external assistance which might not always be available. The to developing countries. Programs or routines used in SAS or other software packages are required to enable the inferential analysis of such surveys as well testing the importance of this added sophistication to the results. It is therefore my intention in this thesis to add to the pool of recommendations concerning methodological techniques of health interview surveys that developing countries could follow in order to have:

1) a more structured and efficient data entry phase;

2) a better and quicker data cleaning phase where necessary;

3) improved techniques to analyze the data from complex samples; and

4) various approaches to simplify existing techniques and ensure the quicker release of results.

It is hoped that these recommendations will lead to an increase in the existing pool of methodological techniques appropriate for household surveys in developing countries, and enhance the capability of local management teams conducting surveys such as the NH&NS.

CHAPTER 2

BACKGROUND TO THE LESOTHO NATIONAL HOUSEHOLD HEALTH AND NUTRITION SURVEY (NH&NS)

2.1 OBJECTIVES

The specific objectives of the NH&NS are :

- To provide national and disagregated statistics for Lesotho on: nutritional status of children under 5 and its determinants; recent illness and injury rates, and disability rates for the whole population; access to health services; utilisation of health services and comments on quality of care and reasons for non utilisation; tuberculosis control; access to and utilisation of maternal services; breast feeding patterns; 24 hour diet profiles for children under 24 months; knowledge of the mechanisms of spread and prevention of AIDS.

- To analyze seasonal trends in morbidity, injury, and nutritional status.

- To measure the contribution of various risk factors (socio-economic, demographic, geographic, occupational factors, practice of breast feeding) on morbidity and nutritional status.

- To evaluate questionnaire instruments designed to improve the accuracy of data on illness incidence and duration, and instruments to analyze the types of morbidity.

2.2 ABOUT THE SURVEY

The survey consisted of four rounds. The first three rounds contained 5000 households each. The fourth round, also a sample size of 5000, was a repetition of the first round, where the same 5000 households were re-interviewed. There was in total therefore a sample size of 15000 households (20000 interviews). This allowed a longitudinal survey design using the first and fourth rounds as well as a cross-sectional design using the first three rounds covering different seasons. Unfortunately the data that was collected and entered onto the personal computers for the first round is unusable due to data processing errors. The loss of this data was mainly due to inefficient procedures that were followed in transferring the data from main frame computers to personal computers. Once the data was transferred, files of the same respondent did not link and as a result large sections of the data were unnecessarily deleted. Unfortunately there were, and still are not, sufficient financial resources available to the Bureau of Statistics to reenter this data. Thus for the purpose of this study it will only be possible to analyze the data cross-sectionally.

2.2.1 THE SAMPLING DESIGN

The NH&NS is a stratified two staged sampling design where clusters or Primary Sampling Units (PSU's) form the first stage sampling units, and the households the second stage.

STRATIFICATION

The stratification is by urban/rural, district, and ecological zone. For administrative purposes, Lesotho is divided into ten districts. Each district has its own developmental committee from the Ministry of Health (MOH), to formulate its plans and policies. From a survey point of view, it is possible to distribute the workload evenly between supervisors and enumerators in the district headquarters towns. It was therefore reasonable and logical to regard districts as an important dimension of study when the strata were defined. Within districts, Lesotho has four agro-ecological zones, viz. Lowland, Foothills, Mountains, and Senqui River Valley. These ecological zones experience significantly different climatic conditions which may affect health and nutrition. This was one of the main reasons for the four zones being the second important dimension when the strata were defined for the rural areas. In the urban areas, the households were ordered by districts only and not by zones, and in fact an implicit stratification based on income levels corresponding to high, middle, and low income groups was achieved. This ensured that every important section of the population was represented in the sample and the strata were more homogenous with respect to the variable(s) or characteristic(s) under consideration than the population as a whole (see Chapter 1), which would have the effect of reducing the standard error of the estimates.

SAMPLING OF THE PRIMARY SAMPLING UNITS (PSU's) AND HOUSEHOLDS

The creation of the PSU's (clusters) was achieved by joining adjacent enumeration areas (EA's) from the 1986 Population Census. Due to travel difficulties in the Mountain and Senqui River valley zones each PSU covered between 200-300 households while in the Lowlands and Foothills the PSU covered 500-600 households. In the urban areas PSU's comprised about 300-400 households where care was taken to ensure that the PSU's were as heterogenous as possible (see Chapter 1). This was achieved by combining EA's that differed most, e.g. by economic level, population density, and remoteness, into a PSU.

Within a stratum the PSU's were selected by probability proportional to size systematic sampling, where the number of households within each PSU served as the size measure. To select the PSU's in a stratum, the following procedure was applied:

a) The sampling interval is calculated by dividing the number of listed households in the stratum by the number of PSU's to be selected from the stratum in the master sample. This interval is represented by the letter I.

b) A random number S between 1 and I is selected, which represents the starting point for the systematic sampling.

Households are chosen from the list whose numbers of order are S, S+I, S+2I, S+3I, etc.

c) Finally every PSU from which a household is selected by this systematic sampling procedure, is included in the master sample.

In total 80 rural PSU's and 31 urban PSU's were selected.

Once the first stage sampling units were selected, the households within the selected PSU's served as the second stage sampling frame. Systematic sampling was adopted to select the PSU's as well as the households.

The number of households selected in each rural stratum was calculated by a predetermined overall sampling fraction of 0.016 (see Chapter 4 for more details) which turned out on average to be 60 households per PSU. In the urban areas a predetermined sample size of 48 households per PSU was decided upon. By selecting the same number of households in each PSU within the same stratum, a self-weighting sample was intended.

The Demography and Social Statistics Division in the Lesotho Government was responsible for the overall conduct and co-ordination of the survey. The Field Operation Division was responsible for the administration of the questionnaire. There were 80 permanent rural and 31 temporary urban interviewers employed, as well as 22 supervisors for both the rural and urban areas, so that each interviewer was assigned to one PSU.
As shown in Table 1.1 the PSU's were stratified by administrative districts, agro-ecological zones, and urban/rural split.

Table 1.1 The number of PSU's and households sampled in each stratum in the NH&NS

District/ zone	Lowland		Foothill		Mountain		SRV		TOTAL	
no. PSU's no. h/holds	r	u	r	u	r	u	r	u	r	u
Butha-Buthe	2 377	1 138	2 344	-	2 365	-	-	-	6 1086	1 138
Leribe	7 1223	4 500	4 670	-	2 335	-	-	-	13 2228	4 500
Berea	6 1034	4 540	3 488	-	-	-	-	-	9 1522	4 540
Maseru	6 961	14 1796	4 658	-	2 330	-	-	-	12 1949	14 1796
Mafeteng	6 1057	2 277	4 658	-	-	-	-	-	10 1715	2 277
Mohale's Hoek	4 769	2 277	2 350	-	2 345	-	2 349	-	10 1813	2 277
Quthing	-	-	-	-	2 130	1 140	4 599	-	6 729	1 140
Qacha's Nek	-	-	-	-	2 269	1 138	2 349	-	4 618	1 138
Mokhotlong	-	-	-	-	4 660	1 137	-	-	4 660	1 137
Thaba-Tseka	-	-	-	-	4 661	1 134	2 257	-	6 918	1 134
Total	31 5421	27 3528	19 3168	0	20 3095	4 549	10 1554	0	80 13238	31 4077

u= urban, r= rural

2.2.2 COVERAGE IN THE NH&NS

A household was defined as a group of people living together sharing eating and catering facilities. A household was regarded as either (i) a one-person household, i.e., one single person who makes provision for his own food without combining efforts with any other person on a regular basis, or (ii) a multi-person household, i.e., two or more persons who live and eat together in the same house, compound or apartment. Domestic servants and other employees formed a separate household if they prepared their own meals. A reserve list of households was drawn up and kept at headquarters, for substitution if any of the originally selected households were not traced. Only the survey field director had the authority to substitute a household.

For coverage within households all household members were listed on the household form of the questionnaire, other forms (on morbidity, maternal care etc.) were completed only for individuals who were living in the household at the time of the interview. A person who had been absent for less than one week from the household of which he/she usually is a member was interviewed if his/her return was expected within two days. The rules adopted for visitors was that they were only included as part of the household if they were present for the previous 7 days. This suggests that they had a close relationship with the host family and probably shared the family's class and other status. It was decided that since migrant workers, who were outside of the country at the time of the survey, do not place a burden on the health services, it was acceptable to exclude them. Work seekers (usually in Maseru) were also not included in the survey since they would have been required to pass a medical examination and would not have wanted to report any ill health. Police and the military were included since they make use of governmental health services.

A check of the coverage in the field took place, which was aimed at measuring the extent to which the enumerator was able to cover all the villages in the area, the selected households in the villages, as well as the persons in the households. During the visits in the field, the supervisor randomly selected 1/10th of the selected households in each cluster under supervision. The supervisor completed the same listing of household members as the enumerators were instructed to do, using a supervisor's check form. The supervisor then compared his/her forms against the enumerators' questionnaires.

2.2.3 TYPE OF INFORMATION COLLECTED

See questionnaire in Appendix E

Information on the household as well as on individuals within these households was collected.

Morbidity

Information was collected on the frequency and type of illness, injury, and permanent disability. Length and severity of illnesses was also investigated.

Utilisation

Of interest to the MOH is what people do when they are ill - where they go for health care and for what reasons. Information concerning preventive care was also collected. There were questions on the use of ante-natal care services, whether they had a trained health worker to attend their delivery and whether they made use of post natal services. The action taken by respondents experiencing a chronic cough was also investigated.

Nutritional Status

Another area of concern is the nutritional status of the children. Data on the weight and height were collected. In order to ascertain the causes of this malnutrition, questions concerning the diet of the child were included, whether it was breast fed and for how long, why breast feeding was stopped, and what other foods the child was fed early on in its life. Questions were asked about the source of income, education levels of the head of the household and the mother, and the number of rooms in the house, in an attempt to measure socio-economic determinants of malnutrition and morbidity generally.

Health Education

Information was collected to establish people's knowledge and awareness of AIDS, and their sources of information.

2.3 MY ROLE IN THE SURVEY

Members of the Bureau of Statistics, (BOS), and the Ministry of Health (MOH), Lesotho, conducted the National Household Health and Nutrition Survey (NH&NS) in Lesotho during 1988 and 1989. Substantial financial support for the survey was provided by Unicef, and the World Bank. Since completing the fieldwork in mid 1989, the data had been edited, coded and entered into personal computers. However data processing and analysis capacity has been severely limited, with the BOS staff being dependent on different short-term consultants who lacked continuity and were obviously not around to deal with the many problems which arise on a day to day basis. An inadequate data entry program proved to be another stumbling block in the survey. There were few efficient validity checks and certain other factors concerning the program accounted for many errors that were identified once the data had been entered. Two years after the NH&NS was conducted, this huge data set had not yet been adequately cleaned, and no analysis or reports had yet been produced. It was at this stage that I became involved with the survey. It was agreed that I would supervise the data cleaning and assist in the writing up of reports about the results. In return I would be permitted to make use of the data set for the purpose of this study.

I spent an entire year cleaning the data. Every three to four weeks I would travel to Maseru and along with three members of the BOS we would try and remove the inconsistencies and errors of the data set. I would remain there for up to five days at a time, when we would review the progress made since my previous visit and then plan new approaches until my following visit. The linking of the various data files from the questionnaire was the major concern during the cleaning process. Data was originally deleted when certain files did not link, and this required us to retrieve the questionnaires from storage and re-enter that data. There were sections that were not completed or incorrectly typed in the first place, which also had to be re-entered. During this period I became aware of the inherent problems of the data set which I identified in the interviewing, the editing, and the data processing phases.

CHAPTER 3

SOURCES OF ERRORS

3.1 INTRODUCTION

Sampling errors occur in surveys because they only cover a part, or sample, of the population. Due to the randomness involved in sampling, the sample may for example include units with a "too high" mean value, thus making the estimator of the mean too high and vice versa. By increasing the sample size, it is possible to reduce the size of the sampling errors. The standard error of the mean, and confidence intervals can be used to assess the probable size of the sampling errors. Verma(1982) defines the sampling error of an estimator as "a measure of its variability under the theoretically possible repetitions of the survey in the absence of non-sampling errors."

Non-sampling errors can be defined as all those errors which are not due to sampling. Non-sampling errors are often a more serious source of errors in surveys than sampling errors. Where sampling errors may cause the estimators to fluctuate randomly around the true value, the non-sampling errors may introduce large systematic deviations between the sample results and the true value. Such systematic or fixed errors over replications are known as bias. A biased sample is therefore not representative of the population from which it is drawn. An example of a biased sample is when a telephone directory is used to sample people in a city. The poorer section of the population who cannot afford a telephone are therefore underrepresented in such a sample. Non-sampling errors are made up of coverage errors, nonresponse errors, measurement and recording errors.

Coverage errors arise from the failure to give some units in the target population any, or too low a chance of being included in the survey, or from including ineligible units in the survey, or from having some target population units appear several times in the frame population.

Nonresponse errors occur when information about households or units of observation that are selected in a sample is not collected or only partially collected. This usually occurs because some persons on the sampling frame cannot be located, or refuse to be interviewed, or parts of the interview are not completed e.g. questions on income.

Measurement errors may arise due to an error in response by the respondent, the effects of the interviewer on the respondent's answers to questions, an error arising from poor wording of the questions, or errors made by the respondent due to their failure to recall certain incidents, or inaccurate measuring instruments such as weighing scales.

Recording errors arise in the survey process due to errors made by the interviewer in filling in the responses to the questions, errors made by the editors when checking, changing and coding the responses, and errors made by the data typist in entering the data into a computer.

In recent times there has been much more emphasis placed on the need for a high quality data set in developing countries. This has also been endorsed by the Demographic and Health Surveys (DHS) program (Croft (1992)). It is therefore necessary to be able to identify and understand these errors so that they can be prevented where possible, or else rectified if already present.

The initial analysis of the NH&NS indicated errors in the results which prevented the early release of the reports for the survey. Since I was primarily involved in the data processing phases once the data had been entered onto a personal computer, I will deal in greater detail with the data entry errors and nonresponses mentioned above, which were easily identifiable during this phase. Measurement errors in the anthropometry section of the survey are also addressed in Chapter 5. The type of errors that appeared will be identified, the data cleaning procedures that were devised will be discussed, and the various strategies for preventing these errors using modern day technology will be considered. General problem areas identified by myself and the survey manager will be addressed. The suggestion made by Kroeger(1985) to reduce the sample size of a national survey in favour of a more intensive small-scale survey (see Chapter 1), will be considered in the context of the NH&NS. Finally general recommendations will be made for preventing similar problems to those experienced in the NH&NS from occurring in similar surveys which are to be conducted in the future.

3.2 THE DATA CLEANING PROCEDURE OF THE NH&NS

For almost a year and a half three members of the Bureau of Statistics (BOS) and myself were involved in cleaning the data. The database was a relational database consisting of nine linked files:- a file on the individuals (A1), a household file (A2), 3 morbidity files C1 (disability file), C2 (injury file), and C3 (illness file), two child nutrition files E1 (breastfeeding file) and E4 (anthropometric data file), a maternal care file D1 (see Appendix E) and an AIDS information file. All the files were linked to the individual file A1 (containing each individual's demographic and biographic data) on the variables: survey round, primary sampling unit (PSU), household number (HHNO), and person number (PN). One of the most serious problems of the data set was the failure of the files to merge on these linking variables. From the above it is evident that many data typing errors occurred, especially on the person number (PN) variable which caused the broken link between the files. Initially there were 984 out of 9900 from the E1 an E4 files that failed to merge with the A file, 891 out of 12825 that failed to merge in the D file, and 8704 out of 71001 that failed to merge in the C file. The following general procedure was followed to clean the data:-

(i) The linking variable PSU was checked for invalid values, compared with valid PSU's values, and corrected. The correction was made by comparing the data on the PC with that of a similar PSU number on the questionnaire. If the data matched the correction was made.

(ii) Tests using the PROC FREQ procedure of SAS revealed obvious data errors such as a value of 'G' appearing for the gender variable, when the only possibilities were 'M' or 'F'. The less obvious errors were then corrected. Examples of these were males appearing in the maternal care file; males appearing in the morbidity section with gynaecological diseases; and individuals over 5 years old appearing in the child nutrition section. Once again the questionnaires containing the abovementioned errors were located and corrected. Once these errors were corrected the errors due to linking problems were addressed. Files were merged using the SAS MERGE statement.

(iii) Each file had a flag variable which was set to 1 in order to indicate the presence of an observation. The process began with the child nutrition files E1 and E4.

(iv) Observations were then listed where countA=1 (i.e. the flag variable of file A) and countE1=. (i.e. countE1 is missing) and countE4=1. This would indicate an entire record of an individual that was missing from file E1 but that had data for file E4 and file A. Frequently an entire PSU's data was missing, but most often it was only an individual's data.

(v) The applicable questionnaires were located from the storage room for re-entry or for checking against the existing data. In the case of systematic errors where an entire PSU's data was missing, that data was re-entered onto the PC. Where individual records were missing, that individual's linking variables were checked against the PC data and that of the questionnaire, and the appropriate corrections were made. In most cases the incorrect PN had been entered (see Section 3.3 below for more details of the types of errors that appeared). This procedure was then repeated where countA=1 and countE1=1 and countE4=. . The same steps as above were applied to the maternal care and morbidity data.

(vi) For those data errors that were due to the interviewer, it was decided to make the invalid entry a nonresponse and along with the other nonresponses to use methods such as subgroup weighting or imputation to compensate for this missing data (see Chapter 4). Since this data cleaning phase took place two years after the survey was conducted it was not feasible to locate the respondent to correct the coding errors.

3.3 THE TYPES OF DATA ENTRY ERRORS IDENTIFIED

Using the procedures described above to cleaning the data, the following errors were observed:

(i) The household numbers on some questionnaires were changed at some stage of the survey, without crossing out the incorrect number. This obviously confused the data typist with the result that on a number of occasions the incorrect identification number for a specific household was entered onto the PC. The person performing the editing did not always correct this error.

(ii) There were many instances where an interviewer wrote an invalid PSU on the questionnaire which made it difficult to identify once the data processing phase had begun. Often the date was written in the box allocated for the PSU.

(iii) The process of call-back for those households where the respondents were not initially at home was not always employed. This was evident by the blank questionnaires with just the PSU number and household number filled in, without any date written down for the call-back.

(iv) There were occasions where a large household required two questionnaires and one of the completed questionnaires from the same household got lost.

(v) One of the major causes for the various files (household file, maternal care, child nutrition etc.) not linking with one another was due to an error in the person number (PN), which was one of the identifying variables linking the various forms to one another. The other linking variables are round, PSU, and household numbers. This error originated in Form A where the PN took on the numbering that was printed on the questionnaire. If for some reason a line was crossed out due to a mistake by the interviewer, a new numbering scheme would have been necessary. However this procedure of changing the numbering on the questionnaire was not always implemented, and the data typist used different PNs for different forms in the questionnaire. More thorough checking procedures should have been implemented.

(vi) Males appeared in the maternal care questionnaire. Males appeared in the morbidity questionnaire with female gynaecological diseases. Sometimes gender being incorrectly recorded was corrected manually by looking at the name or PN error.

(vii) The questionnaire did not have a box or a line allocated for the response of certain questions and as a result the interviewer forgot to record the response.

(viii) Duplicates of the same record appeared. These errors most likely occurred due to power failures during data entry where the data typist had to get back into the middle of the record but instead started again.

(ix) Dummy entries i.e. meaningless data with all the entries equal to zero, appeared. These errors were due to the data entry program. If the data typist added another individual in error he/she would have had to proceed through to the end of the questions for that specific file, entering blanks for all the entries. In many cases the blank observations were never deleted.

(x) Typing errors such as PN=0 or household number (HHNO)=000, or alpha numeric characters appeared in a numerical field e.g. PN=M2 also occurred.

(xi) Invalid PSU's and HHNO's appeared regularly.

(xii) There were many cases where data from the questionnaire did not match at all with the data on the PC. The only explanation that could be found for this discrepancy was that the incorrect PSU, HHNO, or round number was entered.

(xiii) A frustrating aspect for the data typist was that the data entry program did not provide a facility to skip from form to form. If for example there were no children under the age of five years in a specific household and therefore no nutrition form required, the data typist was not able to skip the form but would have to press the enter key for every entry in the form in order to progress to the next form. Besides it being frustrating for the data typist, it was also a time consuming exercise which could have been avoided. The most serious consequence of this aspect of the data entry program was that a blank record would be created that did not merge since it had no valid PN.

(xiv) One of the most serious deficiencies of the data entry program was the procedure for correcting errors. If the data typist made a typing error in a previous line, there was no way to get back to that error immediately and make the correction. The only way to overcome this was to exit the file, reenter it again, locate the error, and make the appropriate correction. In many instances the error was never corrected.

(xv) Entire questionnaires or certain parts of the questionnaire of entire PSU's were never entered even though they were later located during the data cleaning phase.

(xvi) Quite a number of questionnaires were missing. They were either lost in transport of the questionnaires to the Central

Bureau of Statistics or the household was never interviewed. This is an error that took place in the field, and therefore it would not be possible to identify the reasons for it happening at this stage. As Cushing(1990) states "The need to keep questionnaires well organized is apparent after you have collected 6000 to 10000 of them. But when the first few hundred trickle in, the need for organization is less apparent."

(xvii) At an early attempted analysis (before I started working on the data set), those sections of the data causing the mismerges were simply deleted. This proved to be a serious problem especially when observations from the household file were deleted. As a result there was no way of relating morbidity or nutritional status to variables such as education level or occupation. A better approach would have been first to identify the reasons for the error and then make the necessary alterations. In many of these cases, that part of the questionnaire that was missing could have been retrieved from the storage room and then re-entered by the data typist. This data was subsequently reentered. Although the recommendation for correcting the failed merges is obvious, identifying the reasons for them occurring is quite complex, requiring an understanding of the programming language and merging process. Consultants had designed programs and left instructions how to process relational data files, with the instruction to delete those that did not merge, which they anticipated would be very few. There was no consultant present to scrutinise the actual results of each individual merge statement.

(xviii) An entire PSU in the second round was not entered because the field supervisor had observed only after the survey that the interviewer was not conducting any interviews but was instead filling in the questionnaires himself.

Below is an account of the Demographic Health Surveys (DHS) approach to editing missing and inconsistent data and the data processing. Comments are made thereafter of the NH&NS's approach to the editing and processing phases.

3.4 THE DEMOGRAPHIC HEALTH SURVEYS (DHS) EDITING AND PROCESSING STRATEGY (Cushing 1992)

3.4.1 THE DHS EDITING STRATEGY FOR MISSING DATA

(i) Leaving the response blank is an approach that is not encouraged by the DHS, since questions that are not answered due to the nature of the question, are also left blank. This makes it difficult to distinguish between the two approaches.

(ii) Assigning a special code to the missing information will ensure that the missing data will be handled differently in the analysis or imputation phase. It would also serve to indicate how many responses there are if a callback process is considered.

(iii) Deducing a response from the questionnaire needs to be applied carefully, since it could lead to biases. Responses to certain questions can be easily deduced by the flow of questions. It is however a process that is not recommended by the DHS.

The NH&NS's original intention was to distinguish between missing data and data not answered due to the nature of the question by assigning a special code to the missing data only. This approach was unfortunately not consistently executed in either the editing or data entry stages, where many missing data entries were left blank.

3.4.2 THE DHS EDITING STRATEGY FOR INCONSISTENT DATA

(i) Leave the data unchanged.

(ii) Give a special code to indicate the response was inconsistent with other information reported. This will ensure that inconsistent data is not processed, however much of the original information will be lost using this approach. The DHS program advocates the use of a flag variable indicating which response is inconsistent, but leaving the original variable unchanged.

(iii) Deduce a response for the question.

The NH&NS did use flag variables for certain variables e.g. the anthropometric data where a flag variable indicated whether the interviewer was satisfied with the accuracy of the height/weight measurements. An uncooperative child can prevent the interviewer from accurately taking such measurements. However most of the data was entered without such flag variables. A process of cleaning the anthropometric data is considered in Chapter 6 making use of other flag variables at the analysis stage.

3.4.3 THE DHS PROCESSING STRATEGY

The DHS have proposed the following processing strategy (Cushing 1990) :

(i) The development of a self-coding questionnaire where each response has a clear number of digits (usually in the form of the correct number of boxes). This ensures a smoother process during the data entry phase since the questionnaire will not require transcription which is time consuming and is the cause of many errors during data processing.

(ii) The development of an integrated software package (Integrated System for Survey Analysis (ISSA)) which deals with all processing activities, i.e. data entry, verification, editing, file construction and tabulation. This ensures that there is no time wasted in changing the format of the data, for example from the data processing package to the data editing package.

(iii) The use of an intelligent data entry program to test the structure and range of the data, as well as to include the necessary skip instructions. This saves much time in the editing phase where only complex internal checks should be inspected.

(iv) Placing microcomputers in the organization responsible for fielding the survey, in order to process the data. Many developing countries do not have a professional data processing unit and as a result have to employ temporary staff outside of the survey organization to process the data.

(v) Process the data within a two-to three week period rather than wait to edit the data after all the data has been entered. The finalised questionnaire with all the correct codes would need to be available for programming before the fieldwork begins, and as a result it has been suggested that open ended questions should be transferred to a separate file. The data processing can begin after an initial week or two of data entry/editing training. Using three to four microcomputers the data entry phase would be able to keep up with the flow of questionnaires from the field. Many surveys still make use of manual editing which holds up the process. The DHS processing package produces error messages upon which the data entry person has to react. These errors can be tabulated by the survey team, indicating the most evident errors, which would in turn improve data collection procedures.

(vi) Completion of data processing within 2-3 weeks. "As interviewing in each PSU is completed, questionnaires are returned from the field and checked against a field control sheet to make certain that none have been lost. Each PSU is then assigned to a data entry clerk for preliminary manual editing which consists of a household check for eligible respondents, a check of birth histories for correct order and a check of the health sections for correct structure. The questionnaires are then entered into a diskette data file consisting of data from that PSU only. Each PSU requires approximately one day for data entry. After 5 to 7 days of data entry activities, data entry diskettes are collected and data verification begins"

(Cushing(1990)). ISSA is then used to reenter and verify questionnaires that are randomly selected. A secondary editing program then checks that the correct number of questionnaires have been entered and produces a listing of errors for the more complex inconsistencies in the data.

The NH&NS did employ a self-coded questionnaire. Unfortunately some questions which were for routing purposes only (i.e. to indicate to the interviewer where to skip to for the next question) did not have a block assigned (since this data would not have been entered for analysis purposes), yet must have made it confusing for the interviewers, the editors, and the data typists. An example of this error was the question on previous births (question 5 - File D). No box was assigned to this variable which resulted in the interviewer sometimes filling it in or else leaving it blank. Fortunately the editor could check this response against other responses to ascertain what the original response was intended to be.

The data processing of the NH&NS proved to be a major stumbling block. The personal computers initially donated by Sweden used the CPM operating system which could not run SAS or DBASE 3. The database programs that were available on these computers were not known by the local staff. A DBASE 2 data entry program was initially used for data entry, which from Sections 3.2 and 3.3 was clearly inadequate for the job. It was eventually decided to obtain DOS based computers and convert the data to a format compatible with SAS. According to those involved in the data processing of the NH&NS, this process proved to be time consuming, and it is not known whether any data was lost or altered in the process. The SAS editor procedure (PROC FSEDIT) was used during the data cleaning phase.

Interviewers were used as editors and data typists, since there was an interval of approximately three months between each round of the survey during which the interviewer might otherwise have been unemployed or recruited to other jobs. Unfortunately these interviewers were not appropriately trained to perform the functions of a data typist, and since the same team was interviewing, editing and typing, the data entry was necessarily delayed. The final round of the survey was only entered some two years after the fieldwork was completed. This obviously ruled out any attempts to go back to the households or even the interviewers and supervisors to resolve data collection and recording problems.

3.5 SURVEY PROCESSING IN THE FUTURE IN DEVELOPING COUNTRIES

With the development of computer technology, there are options available today to prevent the type of data processing errors that occurred in the NH&NS. Below is a description of some of these options.

Prior to 1986 mainframe computers were used to process survey data in developing countries. On average it took between 2 to 4 years to process the data. Cantor and Rojas (1990) discuss the effect microcomputers have had on survey processing over the past few years. They discuss the gradual upgrading of microcomputers from 286 processors to 386 processors and eventually to 486 processors, associated with a marked increase in processing speed (at times the 386 processor used by the NH&NS took up to half an hour to process one result). 286 processors are currently being used for data entry only. Survey processing software such as ISSA is available, which has an efficient data entry module with range, skip and consistency checks. (Unfortunately not all developing countries have access to such software.) The current area of concern is in the time taken to collect the data and enter it into the computer. Two new approaches encouraged by Cantor and Rojas (1990) for data collection are the use of laptop computers or a notepad sized computer with a pen-based input. Using either of these approaches would mean that an interviewer would not have to venture into the field with a bundle of paper questionnaires on which he/she would fill in the responses. Instead all the responses would be entered directly into the laptop or notepad computer.

3.5.1 THE LAPTOP COMPUTER

An experiment was carried out in Guatemala in 1988 where 300 women were interviewed at two separate times. One of the times a paper questionnaire was used, and the other time the information was recorded directly into a laptop computer data file. Based on this experiment the following advantages and disadvantages of using a laptop computer were observed (Cantor and Rojas 1990).

ADVANTAGES OF THE LAPTOP

(i) The laptop interviews took on average seven minutes less than those done with a paper questionnaire.

(ii) There were slightly fewer error messages for the laptop as opposed to the paper ones that were entered later.

(iii) There is no need for transporting questionnaires, office editing, and keypunching questionnaires at a central office.

(iv) There are no costs for the printing of questionnaires.

(v) If errors are discovered or improvements made in the questionnaire, it would only be necessary to update the input screen with no additional costs for adjustments.

(vi) If programs are written to assess the quality of the data, or check inconsistencies for a single questionnaire, this process could immediately be implemented and the errors could immediately be rectified in the field and would not require a time lapse due to the editing process at the central office.

(vii) With the introduction of the laptop the interviewer takes the role of the data typist, the data editor, as well as the data collector. This would result in a substantial reduction in personnel with resultant huge savings in the cost of the survey. These savings would go a long way to justify the extra money spent on the laptop. Many of the errors identified above in the NH&NS would have been avoided using the laptop. The work burden placed on the survey manager to monitor every stage of the survey would also have been reduced using the laptop.

DISADVANTACES OF THE LAPTOP

(i) Laptops weigh between 1.5 and 3.5 kilograms, which is heavier than a few paper questionnaires.

(ii) An entire set of interviews could be lost due to a bad diskette or disk drive. Backups would need to be made regularly.

(iii) The duration of the battery life for the laptop is limited and this could cause a problem in developing countries where electrical power in the household is generally unavailable. However it has been noted that laptop technology has improved and a laptop can be charged relatively quickly off a car cigarette lighter. However in the Lesotho case, some interviewers would have been interviewing for eight hours a day, for several days a week without access to any power source to recharge batteries.

(iv) The cost of laptops is still quite high, but like microcomputers, prices are coming down with developing technology.

(v) The laptop is not rugged enough for use in rural areas and very dusty conditions. In a trial of handheld computers in Lesotho, the interviews were generally conducted in poorly lit huts and the LCD screens could not be read (M. Price, personal communication). This demonstrated the need for backlit screens, requiring more battery power.

(vi) Most questions have an "open" component for comments which would require reasonable typing skills in order not to slow down the interview.

(vii) An interviewer might be placed at risk (target for theft) in remote or inner city areas of developing countries when carrying expensive equipment.

(viii) If a software or hardware problem occurs in the field, the interviewer would not usually have the skill to resolve the problem.

3.5.2 THE NOTEPAD PEN BASED PC WITH HANDWRITING RECOGNITION

These notepads are set to make an even greater impact on the survey process than laptops. Prototypes have been developed and they weigh 0.5 or 1 kilogram, have a screen matrix of 5 by 8 inches, and they come with no key board, but instead with a pen which allows direct input onto the screen.

ADVANTAGES OF THE NOTEPAD COMPUTER

(i) The notepad computer resembling a writing pad or clipboard is not as threatening to a respondent as a laptop might be.

(ii) With a notepad the interviewer is more mobile and will not have to be seated in a specific place as in the case of a laptop.

(iii) The notepad has a long battery life.

(iv) The handwritten responses of the interviewer are instantly converted to an ASCII text and the data is stored in RAM (random access memory) before downloading takes place. Models from 1996, have between 1 to 3 megabytes of memory available in RAM to store the data. There is therefore no need for diskettes and no risk of problems associated with faulty disk drives.

(vi) Graphically the notepad resembles a paper questionnaire which makes it easier for the interviewer to conceptualize.

DISADVANTAGES OF THE NOTEPAD

(i) At the moment the high costs of instituting a notepad for data collection make it unfeasible financially.

(ii) The notepad is not rugged enough for use in rural areas.

(iii) There are still many flaws with handwriting recognition technology.

3.6 MANAGEMENT IN THE NH&NS - PROBLEM AREAS IDENTIFIED BY THE SURVEY MANAGER

Cushing (1991) had the following to say on the subject of supervision: "Whether the data entry operators are excellent or mediocre, excellent supervision of them is essential. And supervisors should have the support of documented editing rules by which their decisions can be made".

The survey manager of the Lesotho NH&NS had the following comments to make regarding the inherent problems of the survey (Thakhisi 1990).

MONITORING THE SURVEY PROCESS

Thakhisi commented that "there was not effective use of management tools such as work plans e.g. the Ghant chart, coupled with effective monitoring systems and prompt counter acting against the hindering problems. The use of management tools such as schedules assist in the early identification of unplanned errors. Immediate corrective measures are called for, while negligence causes a series of unnecessary hiccups resulting in a failure to meet deadlines."

INVOLVEMENT OF THE STAFF

Concerning the involvement of the various members of staff : "It is important that all key persons be involved at some level of the planning stage so that they fully appreciate the importance of the undertaking. Involvement at a co-ordinating meeting level would be ideal. With the NH&NS the key persons such as field officers, programmers, and printers were never involved, yet their personal commitments were very crucial."

LACK OF TRANSPORT AND DATA PROCESSING EQUIPMENT

"The switch of other BOS surveys from the mainframe computer to personal computers turned out to be a major problem since the NH&NS had to share the data entry equipment with two other surveys which were running concurrently. There was a lot of commotion over who was to use the equipment, and until when. This also required the NH&NS to assign a qualified microcomputer programmer to take charge of the data processing - there was none. Two apprentice microcomputer programmers were assigned to the NH&NS. Without skilled guidance and training, the programmers struggled with minimal progress. The programmers continued for some lengthy periods to work blindfoldedly, without following work plans and target dates.

The lack of a survey vehicle was purely a mismanagement case since the Bureau of Statistics had plenty of vehicles which were available in the 1986 Census."

A serious issue identified by the survey manager was the lack of support he received from the senior management. He felt that many of his complaints about the survey, whether it was problems concerning the fieldwork or programming, were not given enough consideration. It appears from the above excerpts of the survey manager's report that there was a lack of communication between the various members of the survey team. It seems that the survey manager, being a key person in the survey team, was not provided with the necessary supervision and support to ensure the smooth running of a survey of this nature. There were various overseas consultants employed at different stages of the survey with different approaches. This seemed to have a destabilising effect on the smooth running of the survey.

The survey manager was not provided with appropriate training or support to manage the survey effectively. He was not made aware to scrutinise finer details that cause errors in a survey (e.g. lack of validity checks in the data entry program), which is crucial in such a large data set where much can go wrong if these errors are not identified and rectified at an early stage.

3.7 RECOMMENDATIONS FOR FUTURE SURVEYS BASED ON THE NH&NS EXPERIENCE

3.7.1 ERRORS RESULTING FROM POOR QUALITY "IDENTIFIER" INFORMATION

Far stricter supervision and clearer guidelines should have been applied to the data used to identify PSU, survey round, household number and person number - the key data linking all the files. Corrections made to this data should have required personal authorization and checking by a senior supervisor to ensure that they were valid, and that corrections made to one form were carried over consistently to all parts of the same questionnaire and that the corrections were unambiguous to the typist.

3.7.2 PREVENTION OF DATA ENTRY ERRORS

One can deduce from the above discussion of data entry errors that a far more efficient data entry program should have been introduced from the outset with far better validity checks, easier manoeuvring for the data typist between files, and appropriate software to facilitate the linking between the different files. Pre-testing should have been performed on a subset of the data to identify the problematic areas of the entry program. There were several reasons for the poor data entry program. First was the lack of time between completing the final version of the questionnaire (after pretesting and piloting had led to revisions) and starting the fieldwork. Second, was the inherent limitations of the software package chosen for data entry. (The choice, as mentioned above, was restricted by the hardware.) Third was the poor coordination between the foreign consultants - one of whom was contracted to design the questionnaire and another to write the data entry program for several surveys and did not have any involvement in the content of the health survey - hence was not well equipped to design more complex internal validity checks (e.g. ensuring gynaecological illness was not accepted in males). Another reason given for the minimal number of error checks in the data entry programs was so that it would not slow down the entire data entry phase. This proved to be an error of judgement as the resulting number of errors prevented the timely release of any meaningful results from the NH&NS.

For incorrectly typed data which fell in the range of possible responses, another data typist could have been employed to type in the same data, and a program could have been written to compare the entries, so that immediate corrective measures could have been made. I feel the extra cost incurred in employing extra data typists for the job would have been more than justified in the long term.

3.7.3 A COMPUTERISED INTERVIEWING SYSTEM - FEATURES OF THE UNIT

Taking all the abovementioned data capturing problems into consideration I endorse the proposition of the DHS of implementing a computer based interview. Below is a description of the features of a palmtop computer recently released(PSION (1995)). Although this unit is still keyboard driven (In Section 3.5.2 a pen based system was endorsed), it overcomes the disadvantages mentioned above of using a laptop or notepad computer in a developing country and still maintains the advantages of contributing to a quick interview process and a far more accurate and complete data set.

It has the following features:

(i) The unit is lightweight and ergonomically designed, splash proof, and it can be used in a wide range of temperatures.

(ii) The casing is of a rugged design and it can withstand drops and knocks.

(iii) It has a large black and grey scale screen that can display a graphic windows interface for simplicity of operation, and can have backlighting for extra clarity at all times.

(iv) Two penlight batteries provide sufficient power for at least a day of continuous data entry. One therefore has the option to recharge or replace the batteries.

(v) There are two slots for Solid State Disks which enable up to 8 megabytes more memory to be added to the unit.

(vi) The unit's software is IBM-PC compatible.

(vii) The storage system in the unit is upgradeable and cost effective.

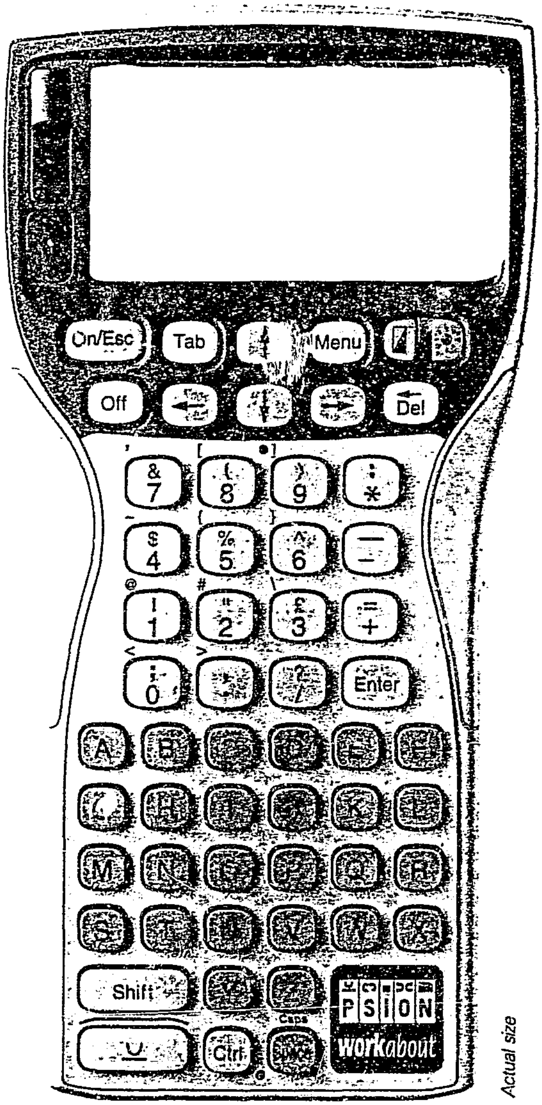
(viii) The unit has a serial port in order to transfer data to a central PC-based database for storage and analysis.

(ix) The unit has the full range of keys for any type of data entry.

(x) The unit's software allows for easy manoeuvrability for the interviewer between questions; appropriate error detection; and easy error correction facilities.

(xi) The unit is cost effective in comparison to laptop or notepads (approximately 60% cheaper than notepads, and 40% cheaper than laptops).

Figure 3.1 The features of a palmtop computer that are recommended for conducting a survey such as the NH&NS



Actual size

Workabout uses an industry standard 16-bit processor in a multitasking operating environment. It has a large black and grey-scale screen that can display a graphic windows interface for complete simplicity of operation, and can have backlighting for extra clarity at all times. With up to one megabyte of RAM memory on board, and Solid State Disks (SSDs) providing up to 8 megabytes more, it can meet any data storage requirement, securely.

A variety of Workabout models is available, providing support for communications and automatic ID. An integral programming language, plus the availability of PC windows based software development tools, makes tailoring the Workabout to precise company application requirements a very easy, quickly achievable task.

All models are lightweight and ergonomically designed, splash-proof (IP54), can be used in a wide range of temperatures, and are rugged enough to withstand being dropped from a height of one metre onto concrete.

Feature rich and user friendly, Psion Workabout offers new scope for improved efficiency in all companies and organisations with staff working away from base.

○ Low cost
○ Feature rich specification - Powerful 16 bit windows Multitasking
○ Easy for all to use
○ Ergonomically styled
○ Programming tools for easy application development
○ Range of peripherals and accessories

A photograph of the PSION and all its features is shown in Figure 3.1.

3.7.4 AN EXAMPLE QUESTIONNAIRE

In order to test the capabilities of the abovementioned palmtop unit, example questions from the NH&NS questionnaire have been chosen and programmed. The overall structure of the NH&NS is maintained by keeping the same linking variables. The example questionnaire and the program source code are found in Appendices F and G respectively.

As in the NH&NS questionnaire there are various files (maternal care, household, and injuries) linked by the variables ROUND, PSU, HOUSEHOLD NUMBER, and PERSON NUMBER. These variables are initially entered by the interviewer in the Demographics file and the program then puts these linking variables into each file to establish a link between the files. By merely placing the marker on the appropriate option, the data is entered into the file. If for example the respondent is male and over the age of 15 years, the next question asked will be that on mining in the mining file. If the respondent is female and over the age of 15 years, the mining question is skipped, and the next question that appears on the screen is from the maternal care section. If the respondent is under the age of 15 years, then neither the mining nor maternal care sections are applicable, and the first question appearing on the screen after the demographic section is from the injuries section. The variables in the mining and maternal care sections are then automatically assigned with the values NA (non applicable). If the respondent is female but not pregnant i.e. Question 6a) = 'No', the next question appearing on the screen will be from the injury section.

If the interviewer makes a typing error, pressing the 'ESC' key takes the interviewer back to the previous screen to make the appropriate correction. The data is only saved once the questionnaire is completed.

Invalid data entries e.g. an option of '8' for Question 7a) (see Appendix F) induces an error message, and the interviewer is asked to retype the correct entry.

Many of the errors in the NH&NS would have been prevented had a similar computerised data capturing unit been used. The problem of linking the various files is immediately solved by this unit. The program guides the interviewer to the appropriate question, and he/she is not confused by arrows or 'GO TO QUESTION' instructions, as can be seen in the NH&NS questionnaire. No invalid data entries are permitted, and as a result the data cleaning phase is almost eliminated. The program is extremely user friendly, and the graphics facilities make data entry very simple for the interviewer.
Once the interview/s is/are completed, the data collected by the interviewer for the day or week is transferred to a central PC by means of a serial port where the analysis will be performed.

It must be stressed though that before introducing a computerised interviewing unit as described above, extensive testing on pilot surveys must be conducted, in order to identify any pitfalls due either to the software, the hardware, or to problems experienced by the interviewer.

3.7.5 SMALLER SCALE SURVEYS FOR DEVELOPING COUNTRIES

The suggestion made by Kroeger(1985) to reduce the sample size of a national survey in favour of a more intensive small-scale survey (see Chapter 1), gives extra weight in a developing country without the appropriate expertise to handle such large-scale surveys. The NH&NS which consisted of four rounds (of which an entire round of data was lost due to mismanagement), would benefit from such a suggestion. The survey could consist of a single round of data (i.e. 5000 households instead of 20000 households), with the following benefits:

- There would be no pressure on editors or data processors to complete their work in time for the next round data. For instance in the NH&NS the same people were used as inteviewers, editors, and data typists.
- Procedures of extensive follow-ups could be implemented in the case of missing or lost data.
- A single round survey makes less demand on consultants' time and would likely lead to greater continuity and fewer consultants being used.
- Financially, the survey would be cheaper.
- The quality of the data collection and processing would be far higher due to more intensive supervision, greater preparation of questionnaires and greater feasibility of double-entry.
- The smaller data set would be quicker to enter and clean, leading to much earlier analysis and reports.
- A smaller, well conducted-survey would instill a sense of confidence in the local staff to carry out similar surveys independently in the future.

There are however certain considerations that need to be made with such a decision. The surveyor has to accept that a smaller scale survey will produce less precise estimates i.e. the confidence intervals of the estimates will be wider. The surveyor will also need to consider the purpose of the survey and what type of data is being collected before such a decision is made. In the NH&NS it can be shown that for specific variables (e.g. proportion of the population over twenty years old with a matric or higher education) the estimates and standard errors of the estimates do not change significantly over the three rounds, i.e. the seasonal change from one round to the next does not have an effect on the proportion of the educated population. From Table 3.1 it is evident that the estimates of the proportion of all the educated in the population from each round falls within the 95% confidence intervals of the other rounds. For example there is a proportion of 0.0483 educated in the June survey data which falls within the 95% confidence intervals (0.0414;0.0610) and (0.0401;0.0550) for October and February respectively (see

Chapter 4 for estimating the proportion and confidence intervals from a complex sample). The confidence intervals for these proportions for the three rounds as well as the combined rounds are plotted in Figure 3.2. The combined round confidence interval is the narrowest with similar interval widths observed for the three rounds, and no real trend. The chi-square test statistic X^2 (see Chapter 6 for analysing two-tables from a complex survey design), indicates there is no significant difference (p=0.5198) between the proportion of educated individuals over the three survey rounds.

Table 3.1 The proportion of educated individuals over the age of 20 years, by round

round	EDUCATED

June 1988	R = 0.0483
	C.I. = (0.0392;0.0573)

October 1988	R = 0.0514
	C.I. = (0.0414;0.0610)

February 1989	R = 0.0476
	C.I. = (0.0401;0.0550)

COMBINED ROUNDS	R = 0.0489
	C.I. = (0.0411;0.0567)

$X^2 = 1.32$, p=0.5198

On the other hand, some variables do have a seasonal pattern which may be important to identify. For example, Table 3.2 and Figure 3.3, show that the estimated proportion (r=0.8740) of children that are still breastfed in October is substantially higher than the proportions still breastfed in June and February, i.e. r=0.8740 does not fall in either of the other round's confidence intervals. A significant difference between the proportion breastfed over the three survey rounds is also indicated (p=0.0031).

Table 3.2 The proportion of children under the age of two years in the rural areas who are still breastfed

round	STILL BREASTFEEDING

June 1988	R = 0.8298
	C.I. = (0.8068;0.8529)

October 1988	R = 0.8740
	C.I. = (0.8530;0.8949)

February 1988	R = 0.8438
	C.I. = (0.8213;0.8663)

COMBINED ROUNDS	R = 0.8492
	C.I. = (0.8337;0.8647)

$X^2 = 12.86$, p=0.0031

This difference in proportions is consistent with Huss-Ashmore and Goodman(1989) who comment that the period of lowest workload for women in Lesotho is from winter to spring (which is around October). They further comment that there is a decreased attention to food preparation and child care (which would include breastfeeding) during high periods of agricultural involvement (such as harvesting) prior to the June months.

One advantage of a multi round survey is the opportunity it provides to learn from problems encountered in the first round, allowing subsequent rounds to achieve a much higher quality of fieldwork and data processing than the first round or a single round survey. This was certainly the experience in the NH&NS.

A multi round survey may be designed to allow for longitudinal follow up at the same households. This adds enormously to the

FIGURE 3.2 Plot of the proportion of educated individuals showing a 95% confidence interval

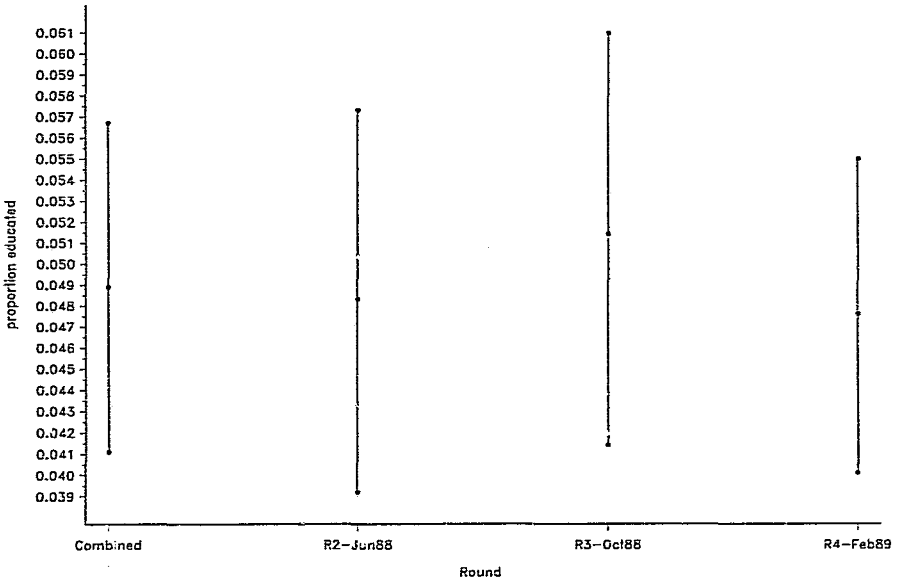
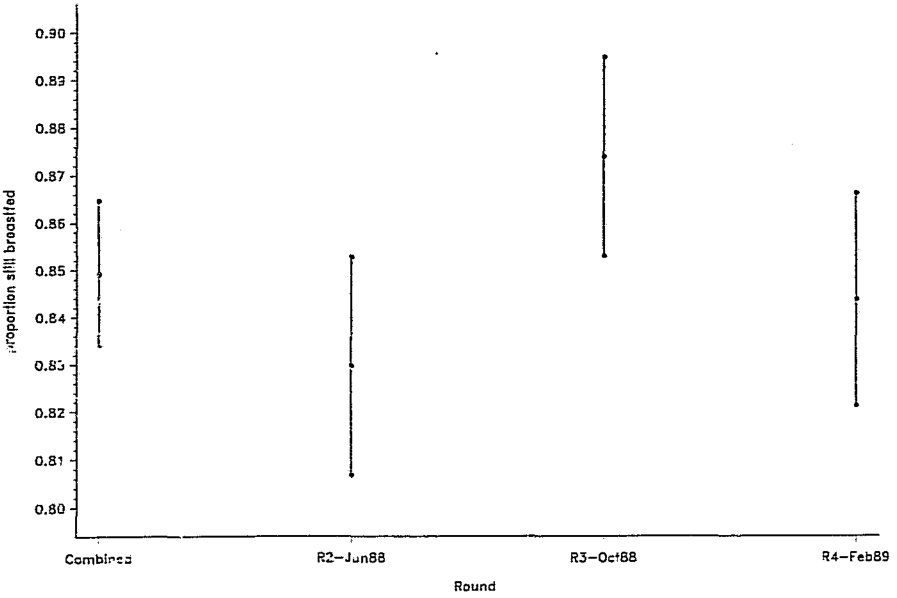


FIGURE 3.3 Plot of the proportion of children under two years who are still breastfed



complexity of a survey - both in the field and in terms of data processing and linkage. However, it also allows for certain unique analyses such as rate of growth, mortality and birth rates, change in household composition, to name a few that were incorporated into the NH&NS.

One of the most unfortunate aspects of the NH&NS was the inability to use all the first round (March 1988) data. Due to serious data processing errors none of that data is available today for analysis. The reason for this was that data was deleted when a mismerge occurred, instead of errors being identified and rectified, which was the procedure followed in rounds 2, 3, and 4. Unfortunately funding to reenter the data at this late stage could not be found. This means that the longitudinal analysis of linking the first round data to the fourth round is no longer a viable consideration of the survey. In my view, the NH&NS was far too ambitious given the capacity available in Lesotho in the late 1980s.

3.7.6 IMPROVING THE MANAGEMENT OF SIMILAR SURVEYS

In a developing country such as Lesotho, where not many surveys of this scale have been conducted in the past, it is essential for the local management team to have constant supervision and guidance from a consultant who is experienced in the field of survey management. This support should also be continuous i.e. the same consultant should be present for the entire duration of the survey. This would prevent confusion among local staff when a new consultant arrives is not fully versed as to the status of the survey at the changeover point, who has a different approach and issues new instructions. It would also ensure that only one consultant would be responsible and accountable for the problems arising, once the survey is completed.

The rounds 2, 3, and 4 data would have lost approximately 10 000 observations had a data cleaning process not been implemented. This comprehensive data cleaning phase occurred in stages: firstly the linkage variables (PSU, round, household, and person number); secondly where individual or household data were missing or incorrectly entered; thirdly where PSU's had large sections blank or were completely missing; fourthly where impossible responses appeared; and lastly where outliers and inconsistencies appeared. For the anthropometric data (see Chapter 6), a data cleaning procedure was also implemented to identify impossible values, outliers, and inconsistencies.

Once this process was completed, compensation for the remaining nonresponses was made by weighting the data at the various levels i.e. at the PSU level, the household level, and finally the individual level. Chapter 4 deals with this process.

CHAPTER 4

WEIGHTING AND COMPENSATING FOR MISSING OBSERVATIONS USING A COMPLEX SURVEY DESIGN

4.1 INTRODUCTION

If each ultimate sampling unit (USU)(which is the household in the NH&NS) has an equal probability of being drawn in the sample, then the sample is called self-weighting, and it is only necessary to weight such a sample if the population totals are required. The standard errors from a self-weighting sample are generally smaller than those from a non-self-weighting sample. If it is necessary to weight the data, and it is observed that the variation between the weights are very small, using weights will have very little effect on the estimated values and the sample data can be treated as if it were from a self weighting sample.

The following types of situation necessitate the weighting of sample data (Stoker 1988):

- If the sample was designed beforehand with a specific purpose in mind, so that the drawn population elements have unequal probabilities of being chosen.

- When a self-weighting sample becomes non-self-weighting as a result of for example varying household sizes - sizes which were unknown during the design of the sample.

- When various subgroups of persons formed by demographic or biographical variables have varying or unequal unit response proportions.

- If there is a non-coverage problem - "When there are differences between the target and sampled populations as a result of which the realized sample, compared with the target population, can be 'skew' in particular demographic and/or biographic variables."

In the NH&NS the data required weighting mainly for the following reasons:

- The totals of various variables for the population was required by the Lesotho Ministry of Health (for example the total number of malnourished children in the population).
- Data was missing for entire households and PSU's (unit nonresponses).
- There were a number of nonresponses for various questions in the survey (item nonresponse).

Stoker(1986) discusses the various implications of using weighted data for statistical analysis:

- Estimators of totals, proportions and percentages which would otherwise be biased, become unbiased.
- The standard errors of the estimators of population characteristics in general increase for the same realized sample size so that the sample design becomes less efficient.
- The various analysis techniques of existing software packages were not designed for weighted data. Many users analysing a survey with a complex design requiring weights, analyze the survey as if it were self-weighting with the result that the estimated standard errors are smaller than they should be, and biased results are produced.

In this chapter the weighting procedure that was applied to the NH&NS is discussed. The procedure starts with the weighting up of the data to the 1986 Census figures, compensating for missing households. Two adjustment factors are included in the final weight: the adjustment factor for the age/sex distribution and a population growth rate factor (since the NH&NS was conducted two-and-a-half years after the Census). Finally a weighting factor to compensate for item non-response is included. Subgroup weighting and imputation are the methods used to compensate for these item nonresponses. By way of example both these approaches will be applied to the data. The abovementioned weighting process is applied separately to each of the three rounds as well as the combined data.

Weighting can be a time consuming and complicated exercise in any survey, and therefore it would be beneficial to a developing country to simplify this phase as far as possible, and at the same time not lose any significant information about the estimates.

It might be possible to exclude the weighting factor to compensate for item nonresponses completely from the final weight. It will be shown that in the case of a complex sampling design, where the PSU is often considered to be the subgroup, the variance of an estimate depends on the variation of the weighting factors of the PSU's within a stratum. The effect of varying weighting factors on the standard error will be investigated in this chapter, by comparing the standard errors of a data set with missing observations, with and without subgroup weighting and with imputation.

Excluding this factor from the overall weight could provide a bureau of statistics in a developing country with a simplified approach to weighting, lessen the workload, and speed up the release of the survey results.

4.2 WEIGHTING IN THE NH&NS

4.2.1 WEIGHTING WITHIN STRATA

The following general procedure was employed to weight the data within the various strata of the NH&NS (see Stoker 1986):

Let P_{h1i} be the probability of drawing the i-th PSU from the h-th stratum and let P_{h2ij} be the probability of drawing the jth USU from the i-th PSU of the h-th stratum given that the i-th PSU has been drawn.

The total probability of drawing the (h,i,j)-th household is then given by

$$P_{hij} = P_{h1i}P_{h2ij} \tag{4.2.1}$$

If sampling takes place proportionally to some measure of size in the first sampling stage, then

$$P_{h1i} = \frac{m_h A_{hi}}{\sum_i A_{hi}} \tag{4.2.2}$$

if m_h PSU's are drawn from the h-th stratum, with A_{hi} the measure of size of the i-th PSU from the h-th stratum. If the i-th PSU contains M_{hi} USU's, of which m_{hi} are drawn with equal probability then

$$P_{h2ij} = \frac{m_{hi}}{M_{hi}} \tag{4.2.3}$$

so that

$$P_{h1i}P_{h2ij} = \frac{m_h A_{hi}}{\sum_i A_{hi}} \cdot \frac{m_{hi}}{M_{hi}} \tag{4.2.4}$$

which simplifies further if all $M_{hi} = A_{hi}$ in which case

$$P_{hij} = P_{h1i}P_{h2ij} = \frac{m_h m_{hi}}{\sum_i A_{hi}} \tag{4.2.5}$$

In order to obtain a self-weighting sample, that is a sample of households drawn in such a way that each household in a particular stratum of the population has an equal total probability of being drawn, one should have

$$P_{hij} = P$$

a constant for all i and j.

In this case for example all $m_{hi} = c_h$ (i.e. an equal number c_h of households are drawn from each drawn PSU in the h-th stratum) so that

$$P_{hij} = P_{h1i}P_{h2ij} = \frac{m_h c_h}{\sum_i A_{hi}} \tag{4.2.6}$$

The weight of the j-th household in the i-th PSU of the h-th stratum is therefore

$$W_{hij} = \frac{1}{P_{hij}} \tag{4.2.7}$$

However, if there are nonresponses in the sampled households then

$$W_{hij} = \frac{1}{r_{hi}\ P_{hij}} \qquad (4.2.8)$$

where r_{hi} is the response rate of the i-th PSU from the h-th stratum.

4.2.2 COMPENSATING FOR MISSING HOUSEHOLDS

Unit nonresponse appears in a survey of individuals when all the information connected to that individual is not collected. This type of missing information occurs as a result of refusals, or when the respondent is not at home at the time of the interview, or call backs e.g. the respondent being away on holiday or in hospital. This type of nonresponse was not a major problem of the NH&NS. There were however entire households where data was not collected. The reasons for this missing data is not entirely clear, but it seems that most of it was due to mismanagement in the field and to data processing errors. Refusals were not commonly reported.

The intended number of households for a specific PSU was therefore not always attained. For the rural stratum of Leribe in the Lowlands zone there are seven PSU's each with 505 households. Based on the 1986 Census there are a total of 27586 households in the population for this stratum. For example for PSU=02107 from this stratum, the intended sample size for each of the three rounds was 63. Considering only the round 2 data of PSU=02107 where only 61 questionnaires were located, equation (4.2.6) becomes

$$r_{hi}P_{hij} \quad = (61/63)x((7x505)/27586)x(63/505)$$

where r_{hi}=61/63, P_{h1i}=(7x505)/27586), and P_{h2ij}=63/505

$$= (7x61)/27586$$

$$= 0.01548$$

The weight for the households of this PSU of this stratum is therefore

$$W_{hij} = 1 / r_{hi}P_{hij}$$
$$= 64.6$$

which indicates that one household from the round two data of the survey for PSU=02107 from the rural, Leribe and Lowland stratum, represents 64.6 households in the population.

For an analysis to be performed on the combined results from all three rounds, a combined weight is necessary. Once again one uses the realised sample size so that, since for PSU=02107 for all three rounds a total of 184 households' data out of an intended total of 3 x 63 = 189 households were collected:

$$P_{hij} = (184/189)x((7x505)/27586)x(189/505)$$

where r_{hi}=184/189

$$= (7x184)/27586$$

$$= 0.04669$$

and the combined weight is

W_{hij} = 21.417 [1]

[1] An external examiner has pointed out that by assuming a sampling rate of 189/505 in this example (which implies that the sample clusters of the three rounds were combined) will result in a much too big estimate of the standard error. The correct procedure would have been to average the estimates over the three rounds and combine their variances.

4.3 ADJUSTING THE WEIGHT FOR DIFFERENCES IN THE AGE-SEX DISTRIBUTION - POPULATION SUBGROUP WEIGHTING

When comparing statistics between separate groups of the population, for example between districts, differences between the districts may depend on differences in the age/sex distribution. Especially in connection with health statistics - for example statistics on morbidity - one often wants to adjust for such differences. The reason is that one wants to analyze whether the differences can be fully explained by the differences in the age/sex distribution or if significant differences in morbidity between the districts remain after "elimination" of the differences in the age/sex distribution. Below a method is described to adjust the statistics for differences in the age/sex distribution, when comparing statistics between groups. This process can also be described as population subgroup weighting, where the various age/sex combinations are regarded as the subgroups. The sample frequencies of these subgroups are then compared to the known population distribution to compensate for the skewness in the data.

4.3.1 THE ADJUSTMENT METHOD

Let us call the survey item x (for example the prevalence of a certain disease). Define the (age/sex) groups to be used for the adjustment. We call them k = 1, ...,s.
The estimate of the total number with the disease for individuals in the age/sex adjustment group k is

$$\sum_h \sum_m W_h X_{hm} 1_{hmk} \qquad (4.3.1)$$

(see Wahlstrom(1990))

where W_h = raising factors or weights for stratum h

X_{hm} = item value for individual m in stratum h:
= 1 if (s)he has the disease
= 0 if not

1_{hmk} = 1 if the individual (h,m) belongs to adjustment group k, otherwise 0

Only 0,1 variables are considered in the following description of the adjustment method, since this type of variable is commonly found in the NH&NS.

The estimated total number with the disease in the population is

$$\sum_k \sum_h \sum_m W_h X_{hm} 1_{hmk} = \sum_h \sum_m W_h X_{hm} \qquad (4.3.2)$$

The estimated total number in the adjustment group k is

$$\sum_h \sum_m W_h 1_{hmk} \qquad (4.3.3)$$

The estimated proportion of individuals in the adjustment group k with the disease X is

$$p_k(x) = \frac{\sum_h \sum_m W_h X_{hm} 1_{hmk}}{\sum_h \sum_m W_h 1_{hmk}} \qquad (4.3.4)$$

The estimated proportion of individuals in adjustment group k in the population is

$$p_k = \frac{\sum_h \sum_m W_h 1_{hmk}}{\sum_k \sum_h \sum_m W_h 1_{hmk}} \tag{4.3.5}$$

where

$$\sum_k \sum_h \sum_m W_h 1_{hmk} = \sum_h \sum_m W_h = N \tag{4.3.6}$$

N = the size of the survey population
Formula (4.3.5) describes the distribution of the proportion of individuals in adjustment groups k=1,...,s in the population.

Therefore

$$\begin{aligned} p_k p_k(X) &= \frac{\sum_h \sum_m W_h X_{hm} 1_{hmk}}{\sum_k \sum_h \sum_m W_h 1_{hmk}} \\ &= \frac{\sum_h \sum_m W_h X_{hm} 1_{hmk}}{N} \end{aligned} \tag{4.3.7}$$

which is the estimated proportion of individuals to have the disease and belong to adjustment group k.

Suppose we want to attach the results to a population where the distribution of the same adjustment groups is q_k where k =1,...,s The corresponding equation to (4.3.7) is then

$$q_k p_k(X) = \frac{q_k \sum_h \sum_m W_h X_{hm} 1_{hmk}}{\sum_h \sum_m W_h 1_{hmk}} \tag{4.3.8}$$

Equation (4.3.8) describes the proportion of individuals to have the disease <u>and</u> belong to the adjustment group k when the distribution of the adjustment groups are q_k.
The estimated total number with the disease in the survey population is

$$\begin{aligned} N\sum_k p_k p_k(X) &= N\sum_k \left(\frac{\sum_h \sum_m W_h X_{hm} 1_{hmk}}{\sum_k \sum_h \sum_m W_h 1_{hmk}}\right) \\ &= \sum_k \sum_h \sum_m W_h X_{hm} 1_{hmk} = \sum_h \sum_m W_h \left[\sum_k 1_{hmk}\right] X_{hm} \end{aligned} \tag{4.3.9}$$

The estimated number with the disease in the adjusted population (still of size N) is

$$\begin{aligned} N\sum_k q_k p_k(X) &= N\sum_k \left(\frac{q_k}{p_k}\right) p_k p_k(X) \\ &= \frac{N\sum_k \left(\frac{q_k}{p_k}\right) \sum_h \sum_m W_h X_{hm} 1_{hmk}}{\sum_k \sum_h \sum_m W_h 1_{hmk}} \end{aligned} \tag{4.3.10}$$

$$= \sum_{h} \sum_{m} W_h [\sum_{k} (\frac{q_k}{p_k}) I_{hmk}] X_{hm} \qquad (4.3.11)$$

i.e. the adjustment factors are (q_k/p_k) $k = 1, \ldots, s$ and the new weights or raising factors to be used in the adjusted population are

$$W_{hijk} = W_{hij} \ (q_k/p_k) \qquad (4.3.12)$$

where $k = 1, \ldots, s$ and W_{hij} is defined in (4.2.7)

4.3.2 APPLYING THE ADJUSTMENT METHOD TO THE NH&NS

The above adjustment procedure (population subgroup weighting) was applied to the NH&NS using the results from the 1986 Population Census to form the ratio q_k/p_k (from (4.3.12)). It is important that the tables from the census and the tables computed from the survey results correspond in order to calculate q_k/p_k. The 'Presence status' from column 4 of Form A of the questionnaire (see Appendix E) is chosen according to which section of the questionnaire is being analyzed. For example only Present=1 (i.e. present for most of last 2 weeks and present now) is chosen in the case of the child nutrition section since the children actually had to be measured. The morbidity section includes the respondents for Pres=1 and Pres=2 (Absent now - present for most of last 2 weeks, will not return in 7 days). Therefore the tables with the same 'present' status as the questionnaire are chosen from the Census. The only tables that correspond to the same breakdown of the 'present' status are the age/sex/district and the age/sex tables. The age/sex/district tables which provide a more detailed breakdown of the population than just the age/sex breakdown, are therefore used.

Using the combined data from all three rounds, the adjustment factor q_k/p_k is the ratio of the age/sex/district totals for the Census (q_k) over the computed survey totals (p_k) taking the weighting factor from (4.2.12) into consideration. Table 4.1 is a frequency table of the adjustment factors of the various cells from the age/sex/district adjustment group. There are altogether 360 cells i.e. 18 age groups for every 5 year age group x 2 sex categories x 10 districts. Only 52% of the adjustment factors lie between 0.85 and 1.15, and the large variation in the adjustment factors might also reflect the age/sex/(rural or urban) breakdown of the population. However since the age/sex/district tables are the only appropriate tables found in the Census tables, they are included in the weighting factor (see equation (4.3.12)).

Table 4.1 Frequency table of the adjustment factors for the age/sex/district adjustment group

Ratio	Frequency	Percent
0.00 - 0.5	3	0.8
0.5 - 0.75	39	10.8
0.75 - 0.85	43	11.9
0.85 - 0.95	60	16.7
0.95 - 1.00	33	9.2
1.00 - 1.05	37	10.3
1.05 - 1.15	55	15.3
1.15 - 1.25	39	10.8
1.25 - 1.50	38	10.6
1.50 - 2.00	9	2.5
2.00 -	4	1.1
Total cells	360	100.0

The same process was applied to each round. However since there is approximately only one third of the combined data in each round, creating 360 cells is problematic since some cells are empty i.e. there are no individuals in a specific age, sex, and district, generated from the survey. The age/sex/district raising factors from the combined rounds are therefore used in each round, since one would not expect this ratio to change dramatically over 6 months, i.e. from the first round to the last round.

4.4 ADJUSTING THE WEIGHTS FOR A POPULATION GROWTH RATE

The population of Lesotho which was 426000 in 1911 increased to 1605177 in 1986 giving an average annual exponential growth rate of 1.8%. This growth rate was not uniform throughout the period. From Table 4.2 it is evident that the growth rate fluctuated between 0.4% in 1936-46 and 2.8% between 1976-86.

Table 4.2 Population size and growth, Lesotho 1911 - 1986

	1911	1921	1936	1946	1956	1966	1976	1986
Population(000)	426	543	661	689	794	968	1217	1605
Annual Growth rate(%)		2.4	1.3	0.4	1.4	2.0	2.3	2.8

The fluctuating growth rates in the period before 1966 could be attributed to various reasons, chief among which are (1) the relative inaccuracy in the data and (2) the varying fertility, mortality and migration. Coming to recent periods, the high growth between 1966 and 1976 is attributed to (1) the under enumeration in the 1966 census, (2) the improvement in enumeration in 1976, (3) the fall in mortality and (4) some increase in fertility between 1966 and 1976 the consequence of improved health, morbidity and mortality. According to the Lesotho Bureau of Statistics the current growth rate of 2.6% is approximately correct, and it is the growth rate which has been adopted by the various users in Lesotho. The 1986 Census was conducted in mid April. The June 1988 round was conducted 2.125 years after the Census, the November 1988 round 2.542 years later, and the March 1989 2.875 years later. Therefore, for example, the growth factor for the June 1988 survey is $e^{0.026x2.125}$.

For the combined rounds the starting date is considered to be between June 1988 and March 1989, which is 2.5 years later, and the growth factor is therefore $e^{0.026x2.5}$.

The weight associated with each observation, taking all the above adjustments into consideration, consists of the following:

WEIGHT1 = WEIGHT (reciprocal of the probability proportional to size adjusted for missing households using the Census data (see Section 4.2))
x AGE/SEX ADJUSTMENT FACTOR (see Section 4.3)
x GROWTH RATE FACTOR (see Section 4.4)
(4.4.1)

The totals for the three rounds as well as the combined rounds are in Table 4.3. The increase of 1407871 from June 1988 to 1423218 November 1988 is due to the growth rate factor.

Table 4.3 Weighted population totals from the NH&NS for 'Pres=1'

June 1988	1407871
November 1988	1423218
March 1989	1435594
Combined Rounds	1421665

4.5 COMPENSATING FOR ITEM NONRESPONSE

This type of missing data appears in a survey when the sampled person is contacted and certain responses to certain questions of the respondent are not obtained by the interviewer. Invalid data are often treated as item nonresponses. The methods of imputation and subgroup weighting for compensating for item nonresponses have been discussed in the literature (Stoker 1985,1997). There are also two approaches that are sometimes followed in practise for compensating for item nonresponses (Stoker 1985,1997) when 50% or more data from a subgroup is missing. The first approach is to combine a subgroup with a weighting factor larger than two (i.e. with a greater than 50% item nonresponse) with a neighbouring subgroup, and thus decrease the number of subgroups. The second approach used by some researchers is to scale all weighting factors larger than 3 to 3. By following this approach, there is the disadvantage of increasing the bias of the estimate, but at the same time there is the advantage of placing a limit on the variation of the weighting factors which would otherwise increase the variance of an estimate. In Section 4.5.3 a resampling procedure is described showing the effect of varying proportions of missing observations on the standard error.

The first step in applying imputation is to identify appropriate subgroups from where the compensation of the missing information will take place. For a complex survey design, the subgroup is usually taken to be a cluster (PSU), since it is expected that respondents in the same cluster are similar i.e. they have similar socioeconomic and geographic characteristics.

However if there is no clustering or if there is reason to believe that more appropriate subgroups than clusters can be formed in the case of a complex survey design, a technique such as CHAID (Chi-Squared Automatic Interaction Detection) analysis (Kass 1980) can be used to identify these subgroups.

This technique is applicable when one has a categorical response variable, which is to be predicted using some or all of a number of available predictor variables. It makes no assumption about the data, and can thus be regarded as a non-parametric technique. The idea behind the technique is that one would like to find groups in the data for which the variability is much lower than in the data set as a whole, by partitioning the data into groups, according to the available predictor variables. One thus examines each predictor variable in turn, and finds the split that gives the greatest reduction in the overall variability of the response variable. The variable that gives the greatest reduction is then chosen as the "splitting" variable. The data are then partitioned into subgroups according to the splitting categories of this variable. Each subgroup is then examined in turn, and further splits made. Each split is only made if it results in a significant reduction (using a Chi-square test) in the variability.

In the case of the NH&NS, the PSU is used as the subgroup in order to apply imputation.

4.5.1 SAMPLE SUBGROUP WEIGHTING

If the p-th subgroup is considered then the raising factor of a particular item in the p-th subgroup is

$$rf(p) = n_p/m_p \qquad (4.5.1)$$

where n_p is the planned sample size and m_p the realized sample size in the p-th subgroup i.e. each item has its own rf(p).
In the case of item nonresponses, a program called ADDFACTOR was written in SAS (see Appendix B) to calculate a raising factor for each PSU (the PSU was considered as the subgroup). For example the raising factors for the first stratum consisting of two PSUs can be seen in Table 4.4 below

Table 4.4 Calculating a raising factor in the case of item nonresponses in Educational Attainment for PSU=01101 and PSU=01108

STRATUM 1	FREQUENCY	PERCENTAGE	RAISING FACTOR
PSU=01101			
VALID	198	95.7	1.045
MISSING	9	4.3	
PSU=01108			
VALID	201	98.0	1.005
MISSING	4	2.0	

Below are the frequencies and percentage before and after subgroup weighting for the variable EDATTAIN (educated/uneducated) from the Round 4 data. Table 4.5 shows that before subgroup weighting 83.2% of respondents from PSU=01101 have less than a Standard 5 education level, whereas after subgroup weighting it is 85.4%.

These raising factors are then incorporated into the final weight so that from (4.4.1)

WEIGHT2 = rf(p) x WEIGHT1 (4.5.2)

Table 4.5 The effect on the frequencies and percentages before and after subgroup weighting, for the education level attained, from the round 4 data.

var=EDATTAIN		FREQUENCY	PERCENTAGE
UNEDUCATED	before	13645	83.2
	after	14010	85.4
EDUCATED	before	2329	14.2
	after	2389	14.6
MISSING	before	425	2.6
	after	0	0
TOTAL		16399	

4.5.2 IMPUTATION FOR ITEM NONRESPONSES

Another method to compensate for item (partial) non-responses in order that a so-called square data set is obtained, is the method of imputation. The advantages associated with the imputation technique are the following:

- it simplifies the statistical analysis of the data set,
- it reduces in general the possible bias in the estimators,
- it leads to consistent results for different analyses, which is not the case in the analysis of a data set with missing observations.

The disadvantages associated with imputation are that:

- it does not guarantee less biased estimators,
- : can increase the size of the standard errors of the

estimators, and thereby reduce their precision.

- an imputed data set can create the impression that the data set is complete, which could result in a false perception of the sample size.

In presenting results of a data set that contains imputed values, it is advisable to analyze the data with and without the imputed values, in order to assess their effect on the precision of the estimators.

IMPUTATION METHODS

Let y_{hi} be the ith value of the item of interest in th3 h-th stratum where h=1,...,g and n_h the number of values in the h-th stratum (this includes l_h missing values). Therefore there are n_h-l_h valid observations.

1) **Global average imputation**
If $\overline{y}_h$ is the average value of y for all respondents that have responded, and y_{mi} is the imputed value of y for the ith respondent where m=1,...,l_h , then it follows that

$$y_{mi} = \overline{y}_h = \frac{1}{n_h - l_h}\sum_i y_{hi}$$

for all missing values of y.

2) **Stochastic global imputation**
In this method

$$y_{mi} = \overline{y}_h + e_{mi}$$

where

$$e_{mi} = y_{hk} - \overline{y}_h$$

and y_{hk} is the y-value of a randomly chosen respondent from the response group of those units which responded for this particular item. Therefore

$$y_{mi} = y_{hk}$$

3) Average value imputation within an imputation class

Imputation classes are formed from subgroups of the data, where households are cross-classified by a few variables such as size of household, or number of people employed, and these are used to create a number of homogenous imputation classes. Therefore a missing observation can be replaced by a valid observation from an imputation class with similar characteristics.

Let $\overline{y}_{hs}$ be the average value of y for the s-th imputation class where all values of y are available, and y_{msi} is the imputed value of y for the i-th respondent of the s-th imputation class where the y-value is missing, then

$$y_{msi} = \overline{y}_{hs}$$

4) Stochastic imputation within an imputation class

The i-th imputed y value in this case is written as

$$y_{msi} = \overline{y}_{hs} + e_{msi}$$

where $e_{msi} = y_{hsk} - \overline{y}_{hs}$

so that $y_{msi} = y_{hsk}$

Here y_{hsk} indicates the y-value of a randomly chosen respondent within the s-th imputation class of those units which responded for this particular item.

5) **The cold-deck procedure**

The cold-deck procedure is applied where households from a previous survey are cross-classified by a few variables such as size of household, or number of people employed, and these are used to create a number of homogenous imputation classes. If in the actual survey there is a nonresponse to a sensitive issue such as income, that information will be taken from the previous survey with the similar classification. To select the missing responses from the imputation class concerned, one can use a random selection of the available responses within that imputation class, or alternatively the average of the responses in the imputation class can also be used (Nordbotten 1963). There are some instances where an adjustment to the imputed value would be necessary. For instance for an item on income, the imputed value obtained from the previous survey would have to be adjusted for inflation which occurred over the time interval between the two surveys.

6) **The hot-deck procedure**

The hot-deck procedure uses an imputed value that is taken from the last preceding unit that falls in the same imputation class from the same survey. The reasoning behind this approach is that it is firstly easy for computer operations and secondly it is expected that a nearby household will be closer to the true value of the nonresponse than the information emanating from a random response from that imputation class. When there is more than one round for a survey, missing information can be obtained from the previous round. Efficient linkage programs would however be necessary.

7) **Stochastic regression imputation method**
The regression model is given here by

$$y_{mi} = \hat{b}_{ro} + \sum_{j} \hat{b}_{rj} z_{ji} + e_{mij}$$

where the z_{ji} are the other responses (which are not missing) that are available from the imputation class from which the particular item is missing, and where the $\hat{b}$'s can be estimated by standard methods such as ordinary least squares.

The assumptions about the distribution of the residual terms are that:

- the residuals are normally distributed with mean zero and equal variances, or that
- the distribution of the residuals e_{mij} is unspecified and are randomly chosen out of the collection of residual values of respondents where y values are present, globally or within an imputation class (Stoker 1985).

4.5.3 AN EXAMPLE ILLUSTRATING THE EFFECT ON THE PRECISION OF AN ESTIMATE FOR VARIOUS PROPORTIONS OF MISSING OBSERVATIONS USING SAMPLE SUBGROUP WEIGHTING AND IMPUTATION

The process of compensating for item nonresponse is illustrated by way of an example using the dichotomous variables, education and gender, from the NH&NS. A person having reached only a Standard 5 education level or lower was considered to be 'uneducated', and any education level above Standard 5 was considered to be 'educated'. Only respondents over the age of 20 years were considered.

In the context of this thesis, a SAS program which we have called RANDALO (random allocation) see (Appendix C) was written to perform the following experiment:

(i) The PSU's within each stratum were randomly assigned to two groups. The PSU was considered to be the subgroup, since it was assumed that people coming from the same PSU's belong to similar socioeconomic groups, and thus have similar education levels.

(ii) Using a SAS randomization procedure, a certain percentage of the data (for a specific variable) in each PSU was assigned to missing (the percentage depended on which group the PSU was assigned to).

(iii) Thereafter the SAS program STANDERR calculated the standard errors of the estimated proportions falling into the four categories in the contingency table (educated females, uneducated females, educated males, uneducated males) obtained from this data set, with and without subgroup weighting, and with imputation. The SAS program IMPUTER (see Appendix D) was written to impute missing observations within a subgroup.

(iv) Steps (i) to (iii) were then repeated 200 times. A data set consisting of the following was therefore created: 200 x 4 standard errors (for the four categories in the contingency table) with subgroup weighting, 200 x 4 standard errors without subgroup weighting, and 200 x 4 standard errors using imputation.

(v) Finally the average standard errors for each of the four cells was calculated for each of the three methods: subgroup weighting, no subgroup weighting, and imputation and these were compared with each other. (Please note the footnote on Page 81)

Experiments 1-3

Following the above procedure, the first experiment investigated a data set with 50% of the observations in each PSU set to missing. This experiment is indicated by 5050 in Table 4.6 i.e. 50% were set to missing in one group and 50% were set to missing in the other group. Similarly Experiment 2 and Experiment 3 were respectively experiments with 10% (indicated by 1010) and 90%

(indicated by 9090) set missing in each PSU. These three experiments represent situations where there is no variation in the proportion missing allocated to each group in (i).

Experiment 4
In the first randomly assigned group (from step (i)) 50% of the observations were randomly set to missing, and in the second group 10% of the observations were set to missing. This experiment represents a data set where there is a big variation in the number of missing observations in the PSU's of each stratum.

Experiment 5
In the first group 10% of the observations were randomly (without replacement) set to missing, and in the second group 90% of the observations were set to missing. This experiment represents a data set where there is an extremely large variation (larger than Experiment 4) in the number of missing observations in the PSU's of each stratum.

Experiment 6
In the first group 90% of the observations were randomly set to missing, and in the second group 50% of the observations were set to missing. This experiment represents a data set where there is a large variation in the number of missing observations in the PSU's of each stratum. The 40% difference in the proportion missing between the two groups is the same as in Experiment 4, however the overall percentage missing is much larger.

The standard errors resulting from these six experiments are given in Table 4.6, where sub=subgroup weighting, nos=no subgroup weighting or imputation, and imp=imputation. There were four factors for each standard error : education level, gender, experiment (proportion missing) and type of weighting technique (sub=subgroup weighting, imp=imputation and nos=no subgroup weighting or imputation). A multifactor analysis of variance procedure was applied to the data to evaluate the effect of each of these factors on the standard error. Of primary interest is the interaction between the factors "weighting" and "experiment", so that a clear indication of the trend of the standard errors for different combinations of subgroup/imputation technique and experiment are shown. The distribution of this data set was slightly skewed to the right and therefore a log transformation was performed on the data which produced a more normally distributed data set. The standard errors are shown in Table 4.6, the means for the different factor levels in Table 4.7 (both for the untransformed and transformed variables), and the results from the analysis of variance procedure are shown in Table 4.8.

Table 4.6 Two-way table of standard errors for six experiments using three approaches to deal with nonresponses

	FEMALES			MALES		
UNEDUCATED	sub	1010	0.0041231	sub	1010	0.0041834
	nos	1010	0.0041231	nos	1010	0.0041834
	imp	1010	0.0041925	imp	1010	0.0042379
	sub	5050	0.0045334	sub	5050	0.0046051
	nos	5050	0.0045334	nos	5050	0.0046051
	imp	5050	0.0047476	imp	5050	0.0048424
	sub	5010	0.0043427	sub	5010	0.0044035
	nos	5010	0.0047173	nos	5010	0.0047617
	imp	5010	0.0044445	imp	5010	0.0045078
	sub	9090	0.0072138	sub	9090	0.0073920
	nos	9090	0.0072138	nos	9090	0.0073920
	imp	9090	0.0074722	imp	9090	0.0076205
	sub	9050	0.0060137	sub	9050	0.0061348
	nos	9050	0.0066900	nos	9050	0.0069118
	imp	9050	0.0062566	imp	9050	0.0063609
	sub	9010	0.0058261	sub	9010	0.0059789
	nos	9010	0.0070551	nos	9010	0.0071103
	imp	9010	0.0059788	imp	9010	0.0061309
EDUCATED	sub	1010	0.0042323	sub	1010	0.0035380
	nos	1010	0.0042323	nos	1010	0.0035380
	imp	1010	0.0042844	imp	1010	0.0035661
	sub	5050	0.0045829	sub	5050	0.0037544
	nos	5050	0.0045829	nos	5050	0.0037544
	imp	5050	0.0047681	imp	5050	0.0038747
	sub	5010	0.0044114	sub	5010	0.0036619
	nos	5010	0.0048398	nos	5010	0.0040435
	imp	5010	0.0045008	imp	5010	0.0037387
	sub	9090	0.0070975	sub	9090	0.0053854
	nos	9090	0.0070975	nos	9090	0.0053854
	imp	9090	0.0072971	imp	9090	0.0055099
	sub	9050	0.0059434	sub	9050	0.0046047
	nos	9050	0.0068065	nos	9050	0.0058671
	imp	9050	0.0061726	imp	9050	0.0047228
	sub	9010	0.0057404	sub	9010	0.0045291
	nos	9010	0.0072338	nos	9010	0.0063203
	imp	9010	0.0058673	imp	9010	0.0046203

Table 4.7 Mean standard error and mean log standard error by gender, education level, subgroup/imputation/none and experiment

```
------------------------------------ GENDER=f --------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        36       0.0055322      0.0011846       0.0041231      0.0074722
     LOGSE     36      -5.2191933      0.2121060      -5.4911500     -4.8965658
     --------------------------------------------------------------------------
------------------------------------ GENDER=m --------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        36       0.0050494      0.0012336       0.0035380      0.0076205
     LOGSE     36      -5.3161086      0.2357399      -5.6441937     -4.8769133
     --------------------------------------------------------------------------
------------------------------------- EDUC=e ---------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        36       0.0050027      0.0011614       0.0035380      0.0072971
     LOGSE     36      -5.3229158      0.2253203      -5.6441937     -4.9202783
     --------------------------------------------------------------------------
------------------------------------- EDUC=u ---------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        36       0.0055789      0.0012351       0.0041231      0.0076205
     LOGSE     36      -5.2123861      0.2198649      -5.4911500     -4.8769133
     --------------------------------------------------------------------------
---------------------------------- SUBIMP=imp --------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        24       0.0052381      0.0011937       0.0035661      0.0076205
     LOGSE     24      -5.2756906      0.2214444      -5.6362827     -4.8769133
     --------------------------------------------------------------------------
---------------------------------- SUBIMP=nos --------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        24       0.0055412      0.0013412       0.0035380      0.0073920
     LOGSE     24      -5.2243756      0.2468705      -5.6441937     -4.9073569
     --------------------------------------------------------------------------
---------------------------------- SUBIMP=sub --------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        24       0.0050930      0.0011406       0.0035380      0.0073920
     LOGSE     24      -5.3028867      0.2169802      -5.6441937     -4.9073569
     --------------------------------------------------------------------------
--------------------------------- EXPERIM=1010 -------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        12       0.0040362    0.000298400       0.0035380      0.0042844
     LOGSE     12      -5.5150976      0.0770596      -5.6441937     -5.4527748
     --------------------------------------------------------------------------
--------------------------------- EXPERIM=5010 -------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        12       0.0043645    0.000374914       0.0036619      0.0048398
     LOGSE     12      -5.4378082      0.0890844      -5.6097731     -5.3303819
     --------------------------------------------------------------------------
--------------------------------- EXPERIM=5050 -------------------------------
     Variable   N            Mean        Std Dev         Minimum        Maximum
     --------------------------------------------------------------------------
     SE        12       0.0044320    0.000397251       0.0037544      0.0048424
     LOGSE     12      -5.4228169      0.0939316      -5.5848268     -5.3303448
     --------------------------------------------------------------------------
```

```
------------------------------------ EXPERIM=9010 -----------------------------------
      Variable   N         Mean      Std Dev        Minimum       Maximum
      --------------------------------------------------------------------
      SE        12    0.0060318  0.000859945      0.0045291     0.0072238
      LOGSE     12   -5.1205473    0.1486801     -5.3972320    -4.9303741
      --------------------------------------------------------------------
------------------------------------ EXPERIM=9050 -----------------------------------
      Variable   N         Mean      Std Dev        Minimum       Maximum
      --------------------------------------------------------------------
      SE        12    0.0060404  0.000724979      0.0046047     0.0069118
      LOGSE     12   -5.1165230    0.1286908     -5.3806778    -4.9745252
      --------------------------------------------------------------------
------------------------------------ EXPERIM=9090 -----------------------------------
      Variable   N         Mean      Std Dev        Minimum       Maximum
      --------------------------------------------------------------------
      SE        12    0.0068398  0.000865509      0.0053854     0.0076205
      LOGSE     12   -4.9931126    0.1363197     -5.2240637    -4.8769133
      --------------------------------------------------------------------
```

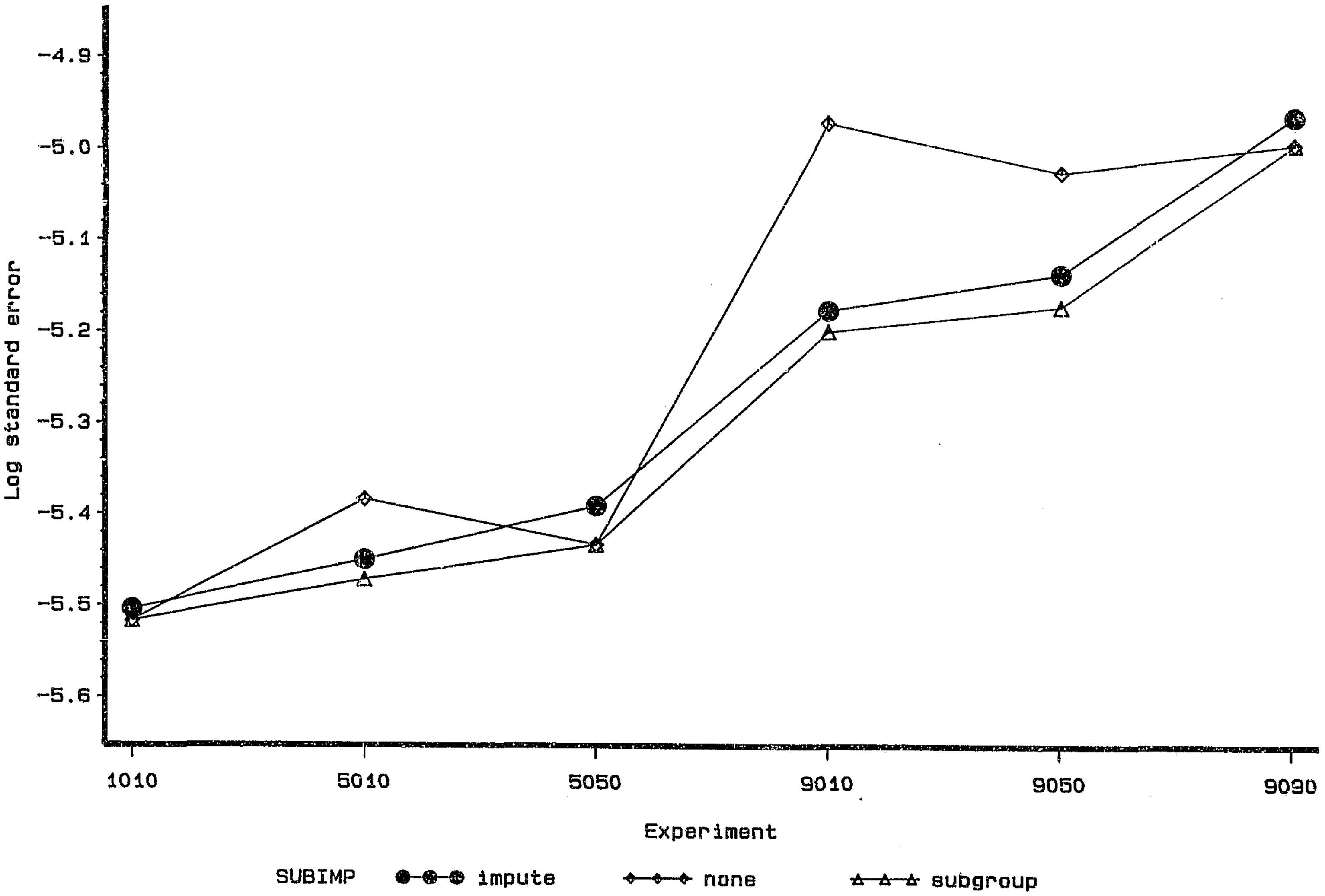
Table 4.8 Analysis of Variance (ANOVA) table of the main effects and the interaction term between 'experiment' and 'subgroup/impute' for Log(SE)

Source	DF	Sum of Squares	Mean Square	F value	Pr > F
GENDER	1	0.16906667	0.16906667	30.77	0.0001
EDUCATION	1	0.21990265	0.21990265	40.02	0.0001
SUBIMP	2	0.07629488	0.03814744	6.94	0.0021
EXPERIMENT	5	2.80932303	0.56186461	102.27	0.0001
SUBIMP*EXPERIMENT	10	0.12846251	0.01284625	2.34	0.0231
Error	52	0.28569597	0.00549415		
Corrected Total	71	3.68874572			

R-square=0.922549

From Tables 4.7 and 4.8 it can be seen that there is a significant difference between the mean log standard errors of males and females (p=0.0001) and of educated and uneducated individuals (p=0.0001) in the study. The significant differences in the mean log standard errors can be attributed to the different sample sizes in the education and gender categories of the NH&NS. Of primary interest is the effect of the type of experiment and the subgroup/imputation effect on the standard error. As main effects there are significant differences between the six experiments (p=0.0001) and between the three approaches in dealing with missing data (p=0.0001). The interaction term SUBIMP*EXPERIMENT is significant (p=0.0108) and an interaction plot is displayed in Figure 4.1.

FIGURE 4.1 Plot of the log standard error reflecting the "subgroup/impute/none" by "experiment" interaction



From Figure 4.1, as is expected, there is seen to be a gradual increase in the mean log standard error (for subgroup weighting, imputation and none) from the 1010 experiment, where the least number of observations were set to missing, to the 9090[2] experiment where the most observations were set to missing. When a constant number of observations were set to missing in both groups (i.e. experiments 1010, 5050 and 9090) the mean log(SE) for SUBIMP=none and SUBIMP=subgroup are the same. This is because the weighting factor for SUBIMP=subgroup is consistently larger (by the same constant) than SUBIMP=none and as a result it has no effect on the standard error. In Chapter 6 the formula for a standard error for a complex survey design is given (see equation 6.2.11). It can be shown that when each observation in the sample has the same weighting factor to compensate for item non-response, the log standard error which is calculated for subgroup weighting is the same as that when no subgroup weight is used.

In each experiment, subgroup weighting performs better than imputation i.e. subgroup weighting has a lower mean log standard error than imputation. The largest differences in log standard errors between **no** subgroup weighting and subgroup/imputation appears in all four groups of experiment 9010 in which there is the largest variation of missing data between the two groups (experiment 5). To a lesser degree this difference is also observed in experiment 9050 (experiment 6).

These results clearly indicate that the larger the **variation** in the proportion of missing data between PSU's in a complex design such as the NH&NS, the more important subgroup weighting or imputation becomes. Based on the above experiment, sample subgroup weighting is recommended in preference over imputation for the following reasons:

[2] An external examiner has stressed that compensation when 90% of the values of a variable is missing is not recommended. However, this does not exclude imputation based on population models.

1) Sample subgroup weighting is less complicated than imputation to implement.

2) Sample subgroup weighting produces more precise estimates (i.e. lower standard errors) than imputation[3].

Improved training and monitoring of the interviewers would go a long way to eliminate the need for subgroup weighting. Highly skilled interviewers conduct an interview efficiently and check their questionnaires thoroughly, thus preventing many item nonresponses and the need for subgroup weighting.

For a bureau of statistics deciding on whether to include or exclude a weighting factor to compensate for item nonresponse, it is necessary to compare the proportions of missing data from the various PSU's within a stratum in order to establish the variation. Where there is only a small variation in the weighting factors due to item nonresponses amongst PSU's in a stratum, those weighting factors can be excluded from the overall weight. However, when there are large variations in the weighting factors, there may be a substantial increase in the magnitude of the standard error if subgroup weighting or imputation are not used. While it would save much time always to exclude subgroup weighting from the weighting procedure, such an approach should be followed with caution i.e. only if item nonresponse is small or if it is evenly spread across PSU's.

[3] An external examiner has indicated that the procedure of calculating standard errors including the imputed values of variables is incorrect. In imputation new real data is not obtained and imputation does not increase the degrees of freedom. The standard errors should be calculated excluding the imputed data.

CHAPTER 5
DEALING WITH ANTHROPOMETRIC DATA

5.1 INTRODUCTION

In developing countries protein energy malnutrition (PEM) has been identified as one of the most important nutritional problems in the world today. In the under 5 year age group, a large proportion of the morbidity and mortality in these countries are attributed to nutritional deficiencies (Kenya (1991)). Due to the importance of nutritional status as an indicator of health status in a population, it has been decided to apply the techniques of the previous chapters to the anthropometric data in the NH&NS. The various issues unique to child nutrition data that can cause errors in the results (such as the choice of cut-off points and using a problematic reference population) are considered. Errors resulting from the data collection phase are explained, as is the data cleaning procedure implemented in the NH&NS. The analysis of the remaining sections of the NH&NS are presented in Chapter 9. PEM is usually characterized by the failure to grow and a deficiency in weight. In anthropometry, in order to assess the nutritional status of children, height and weight are measured and used to construct the following indices: weight-for-age, height-for-age, and weight-for-height. The nutritional status of children has been proposed as a measure of the development of a country. After growth faltering in the first few years of life there is rarely compensatory catch-up growth later on (Sommerfelt 1991).

The World Health Organization (WHO Working group (1986)) identifies inadequate food intake and infections as the primary cause of growth impairment in children in developing countries. They further comment that "a deficit in growth is not necessarily the most sensitive indicator of inadequate nutrition; for example, a marginally inadequate energy intake may cause a reduction in physical activity before there is any impairment in growth.". One has to therefore look at a combination of information when assessing the nutritional status of an individual or population. In addition to anthropometric methods, the following methods to assess nutritional status can also be used:

(i) Dietary methods, where one looks at the inadequacies of one or more nutrients due to low levels in the diet.

(ii) Investigation of certain factors such as the use of certain drugs, dietary components, or disease states, which interfere with the ingestion, absorption, transport, utilization, or excretion of the nutrients.

(iii) Laboratory methods where one would examine a person by using biochemical tests, and/or tests that measure physiological or behavioural functions.

(iv) Clinical methods: i.e. taking a medical history and making a physical examination.

Depending on the study objectives, the costs and the time available, any of the above, could be used to assess the nutritional status of an individual or population.

In this thesis only anthropometric methods are investigated, since it was the only approach used by the BOS in the NH&NS.

Keller (1991) indicates why anthropometric assessments of nutritional assessment using weight and stature should be confined to populations rather than individuals - "Both stature and weight as well as growth in general are very unspecific indicators of nutritional status. To allow a diagnosis of malnutrition in an individual child, the finding of reduced weight or height increments must be accompanied by other indications of malnutrition. For nutritional assessments the measurement of stature and weight alone is therefore mostly applied to populations rather than to individuals. In populations epidemiological and probability considerations can be used to arrive at a nutritional diagnosis if non-nutritional factors are excluded as major causes."

The use of anthropometric data in population studies has several purposes. (1) They can provide a general indication of the magnitude of the problem of malnutrition in the population in order to assist political sensitization and mobilization of resources; (2) they can provide an accurate estimate of the prevalence of malnutrition on a one-time basis and for monitoring trends over time; (3) they can be used to assess the distribution of PEM across various geographic and socioeconomic groups; (4) they can contribute to an analysis of determinants of PEM. (Pelletier (1991)); and (5) they are important in evaluating the impact of intervention programmes such as supplementary feeding programmes.

5.2 USING REFERENCE POPULATIONS

5.2.1 THE REQUIREMENTS OF A REFERENCE POPULATION

A generally healthy, well-nourished child population of sufficient size is the requirement for a reference population (Waterlow et al. 1977)). Keller (1991) gives the following recommendations for a reference population : "In fact, if the reference population is based on data from a generally healthy and well-nourished population, and if there is an equal genetic growth potential in different populations, there seems to be no major reason not to use, with some reservations, a reference population as a norm, representing the growth that could be attained by any child population growing up under similarly favourable conditions.". An international reference population will permit comparisons and analyses of survey results made at different times or in different countries (Graitcer and Gentry (1981)).

Cameron (1991) gives the following reasons for not being able to create adequate reference data in developing countries: "the logistical basis to achieve such standards is simply not available in these countries because they are developing countries without the infrastructure necessary to succeed with these tasks. In short, the nutritional field worker, or indeed the human scientist, who attempts to document the nutritional status of the individual or community in a developing country is constantly faced with less than ideal circumstances within which he has little or no opportunity to exercise 'ideal' procedures".

There has been much debate on the appropriateness of using a healthy, well-nourished population as a reference for a developing country. Pelletier (1991) remarks on the effect of genetic differences when he writes "genetic differences in early child growth are commonly perceived to exist, even though the scientific evidence suggests otherwise." Pelletier(1991) further comments that in most cases, if one were to use a sample of the more privileged children from a developing country, the mean stature would fall between the 35th and 65th percentile of the National Center for Health Statistics (NCHS) reference population. Pelletier finds however the exception to these findings in the Asian region (Japan, Taiwan, Hong Kong) where real genetic differences are found, and a revised reference population is required. Pelletier (1991) also contrasts the means for privileged samples with those of less priveledged samples from the same population and shows that the differences associated with socioeconomic status are far greater than those attributable to genetic differences. Keller(1991) points out that the elites may be of different ethnic origin from the general population. While this has occasionally been found to be true, the overwhelming number of studies in the literature contradict this idea. Studies often discussed in this respect are those of Watson and Lowry (1975) and Wingerd et al. (1971) which show that American black children demonstrate no significant differences in their height attainment to similar white children, proving the equality of genetic potential between African children (who originate from mainly developing countries) and those of other racial groups. Pelletier (1991) suggests that a subsample be chosen of the priveledged section of the population, and a comparison made of the mean stature with that of the reference population, in order to test the applicability of the reference population.

For those cases where "privileged" or "elite" children cannot be easily identified, Cameron(1991) gives a solution by arguing his interpretation of normal growth: "Normal growth must be that reflected by children who live in a particular environment and demonstrate good health. In the context of a developing country 'good health' is usually taken as freedom from illness or disease as reflected in a normal growth rate, i.e. a growth rate that maintains a child's centile position relative to his peers. The fact that the distance centile resulting from this velocity may be below NCHS centiles (see Section 5.2.2), does not, in itself, indicate that the child is unhealthy. It does mean however, that given the environment within which the child is living, the child is managing to maintain a normal growth rate i.e. a growth rate which is not causing it to fall behind its peers. Of major importance is the fact that we do not accept such a growth rate as being equivalent to the target growth which could be achieved given better nutritional and socioeconomic conditions."

Since no better reference population than the NCHS/CDC population exists and because of the fact that it is recommended by the World Health Organization, Z-scores are calculated from this reference population to determine the nutritional status of the NH&NS.

5.2.2 THE NCHS/CDC REFERENCE POPULATION

While reference populations have been established in a number of countries, the World Health Organization in 1978 identified and recommended that the reference population constructed by the combined task force of the National Center for Health Statistics

(NCHS) and the Centers for Disease Control(CDC) was the most suitable for development as the international growth reference. The NCHS/CDC reference population comprises data from the Fels Research Institute (see below) and the Health Examination Surveys (HES) of the NCHS.

The Fels Research Institute data is from a longitudinal study of growth. A convenience sample of 867 children from white middle-class families living near Yellow Springs, Ohio was used. On average 720 children were measured at birth and at age 1, 3, 6, 9, 12, 24, 30, and 36 months between the years 1929 and 1975, and smoothed observed percentile curves of weight-for-age, length-for-age, and weight-for-length were derived (Hamil PVV et al. (1979)).

The NCHS data is from three HES surveys conducted across the United States of America, that used stratified probability sampling designs: HES Cycle I for ages 6-11 years (1963-65), HES Cycle II for ages 12-17 years (1966-70), and the first National Health and Nutrition Examination Survey (HANES I) for ages 2-17 years (1971-74). For this sample smoothed observed percentile curves of weight-for-age, stature-for-age, and weight-for-stature were derived for ages 2-18 years. Due to the design and the location of the Fels Research Institute sample, it cannot be regarded as representative of children in this age range throughout the United States. It is however the best available data set for children under two years. The NCHS data on the other hand can be regarded as reliable population estimates of the attained growth of children in the US for these ages (Dibley et al. 1987).

5.2.3 ERRORS ARISING DUE TO THE NCHS/CDC REFERENCE POPULATION

One of the major problems of using the NCHS/CDC reference population is the discontinuity that was found at the junction point (i.e. at 24 months) of the two distinct data sets from which the reference curves are derived. The discontinuity is observed for all three indices. These discontinuities are caused for the following two reasons:

(i) The Fels and the NCHS data are from two completely different populations. The Fels data is from an unrepresentative convenience sample of middle-class Americans, whereas the NCHS data is from a more representative probability sample.

(ii) The Fels data only has recumbent length measurements, whilst the NCHS reference population has an unknown proportion of recumbent length measurements among the stature measurements between 24 and 36 months. There is no way of knowing which individuals in the NCHS population are measured lying down and which are measured standing up. (It should be noted that length refers to measurements made with the child lying down, while stature refers to measurements made with the child standing.).

Pelletier(1991) describes these issues: "The combining of these two samples results in an artifactual disjunction in medians, percentiles and Z-score cut-off points for weight and height at 24 months of age (with the Fels children showing higher medians and lower variances), which is partially exacerbated by the fact that the NCHS data represents a mixture of lying and standing measurements for stature (height). The net result, is the appearance of a sharp drop in the prevalence of stunting and wasting after 24 months of age, which is most pronounced in the more malnourished populations and is largely an artefact of the mixed standard. This drop has been widely observed in surveys from developing countries, and is also reflected in an increase in mean levels of the various indicators just after 24 months."

The effect of this discontinuity on the prevalence of malnutrition is illustrated by Dibley et al. (1987). One could clearly observe a drop in the prevalence of malnutrition, using height-for-age as an indicator (i.e. percentage below -2 Z-scores) when comparisons of national nutrition surveys in Yemen in 1979, in Swaziland in 1984, and in the United States 1976-1980 were made. The drop in prevalence rates is more pronounced in the developing countries of Yemen and Swaziland. This discontinuity has the effect of falsely indicating to the health authorities that public health resources should be concentrated on the 12-24 month age group.

To compensate for the height issue discussed in 5.2.3(ii) above, it has been suggested that 1.5 cm be subtracted from the Fels data since it is measuring length which is supposed on average to be approximately 1.5cm longer than stature. The problem with using the 1.5cm adjustment is the unknown proportion of recumbent length measurements among the stature measurements between 24 and 36 months for the NCHS population.

The solution provided by Dibley et al. (1987b) is to continue to use the NCHS/CDC reference population until an improved reference is available. In addition it is recommended that data are tabulated and compared separately for age groups below and above 24 months.

Pelletier et al. (1990), Cogill(1982), Yambi(1988), and Briend et al. (1989) investigated the possibility that the distortion in individual Z-scores caused by the drop in prevalence rates discussed above, reduced the power to detect associations with socioeconomic variables. In all the findings of these studies the conclusion was reached that the power to detect associations with socioeconomic variables was not reduced. Pelletier(1991) points out that nutritional monitoring at the population level is unlikely to be affected. However as pointed out by Dibley et al. (1987b) longitudinal surveys and program evaluations may be affected by these artefacts.

Based on the above findings and recommendations, the cautious approach would be to follow Dibleys advice, and analyze the data separately for children below and above 24 months. If the alternative advice is followed, where the artefact is ignored, it may be worth checking for distortions in the results.

5.3 MEASURING ANTHROPOMETRIC INDICES

The anthropometric indices weight-for-age, height-for-age, and weight-for-height can be compared using percentiles, percentage of the median, or Z-scores, based on the NCHS/CDC reference population.

5.3.1 PERCENTILES

A percentile measure refers to the position of the measurement value in relation to all of the measurements for the reference population. If the data is from a normal distribution (e.g. height-for-age) the 50th percentile will correspond to the mean of that distribution. The distribution for weight-for-age and weight-for-height is however skewed, and therefore the 50th percentile will be lower than the mean. Individuals are classified as being 'at risk' of malnutrition if their percentile scores are below the 5th percentile.

The NCHS/CDC data set in its original form consisted of these smoothed percentile curves. Using percentiles to evaluate anthropometric indices in developing countries, has not been recommended by Hamill et al. (1979) and Waterlow et al (1977). The reason for this is that the NCHS reference data are from healthy children in an industrialized country, and there are many populations in less developed countries where large numbers of children are below the extreme percentile of the reference population (i.e. the 5th percentile). This has the effect of making it difficult to accurately classify large numbers of individuals.

5.3.2 PERCENTAGE OF THE MEDIAN

The second approach used to evaluate anthropometric indices, is the percentage of the median. Those individuals below a cut-off point of 80% of the median, have been indicated to represent the malnourished section of the population. The main problem with this approach is that a given percent of median for an indicator at different ages does not have the same meaning for different indicators, and therefore also has different clinical significance.

5.3.3 Z-SCORES

The Z-score has been recommended by Waterlow et al. (1977) to overcome the problems associated with expressing anthropometric indicators as a percentage of the reference median. This method allows for the extrapolation of anthropometric cutoff points beyond the outer percentiles of the original reference data. To permit the expression of growth in terms of standard deviations, the NCHS/CDC percentile curves had to be transformed into a Z-score representation of normalized growth curves. The Z-score is calculated using the following formula:

$$Z\text{-}score = \frac{Individual's\ value - median\ value\ of\ ref\ population}{Standard\ deviation\ value\ of\ ref\ population}$$

A comparison of the means and medians of the reference weight-for-age and weight-for-height distributions reveals a positive skewness, i.e. the means generally exceed the medians, and also that the positive skewness increases with older ages (Dibley 1987a). Due to this skewness in the reference population for these indices, the median was used instead of the mean in calculating the Z-score.

In order to develop normalized growth curves the following process was implemented by Waterlow et al. (1977):

The weight-for-age and weight-for-height reference data were split into upper and lower segments at the median, so that the data would behave more like normal distributions. Each segment was treated as one half of a normal distribution and separate age or height-specific standard deviations were defined for the upper or lower segments of the distribution. The upper standard deviations were the average of the estimated standard deviations from the 75th, 90th, and 95th observed percentiles while the lower standard deviations were the average of the estimated standard deviations from the 25th, 10th, and 5th observed percentiles.

There were very small departures from normality for the height-for-age data which were not sufficiently large enough to warrant the separation of the data into upper and lower segments. For this data a single set of age-specific standard deviations was defined by averaging the estimated standard deviations from the six observed percentile curves i.e. the 5th, 10th, 25th, 75th, 90th, and 95th percentiles.

The standard deviations estimated from each observed percentile were calculated by taking the absolute deviation from the median for the observed percentile at each month of age or centimetre of height and dividing it by the Z-score value that would coincide with that percentile.

Dibley et al. (1987a) comments that using separate upper and lower standard deviations for the weight-for-age and weight-for-height curves, represents a crude transformation and that a log transformation may have had more desirable statistical properties. They point out, though, that the disadvantage of using log weight-for-age and log weight-for-height curves is that this would require nutrition field workers to transform data before plotting points on growth charts. It was further commented that the possibility of a log road-to-health chart seemed extremely impractical and, therefore this approach was not adopted.

Curves were fitted to each of these sets of calculated normalized percentile values by polynomial regression and cubic splining techniques. The complete tables of the normalized growth reference curves with weight and height values for selected percentiles and Z-scores have been published (Waterlow et al. (1977)), WHO(1979)).

This method allows anthropometric indicators to be defined by extrapolation beyond the observed outer percentiles of the original reference data. Using Z-scores has the advantage in that they can be accurately defined in the lower extremes. This would ensure the classification of all individuals, which is especially necessary when studying a malnourished population. Based on this reasoning, Waterlow et al. (1977) recommend that Z-scores be used in preference over percentages of the median where large numbers of children fall above the upper or below the lower centiles. Z-scores are therefore used to assess malnutrition in the NH&NS, since Lesotho is a developing country where one would expect a large proportion of malnourished individuals to fall in the lower extremes of the population.

5.4 THE CHOICE OF AN ANTHROPOMETRIC INDEX

The **height-for-age** index is an indicator of linear growth retardation. Children whose height-for-age is below minus two standard deviations (-2SD) from the median of the reference population (or less than 3 percentiles of the reference population) are considered short for their age. They are also referred to as "stunted" and are considered to be chronically undernourished. There seems to be a direct relationship between stunting and poor socioeconomic status which gives rise to inadequate living conditions not conducive to attaining optimal health. It is also frequently associated with chronic or repeated infections.

The **weight-for-height** index measures body mass in relation to body length, and describes current nutritional status. Children below (-2SD) from the median of the reference population are considered thin (also referred to as "wasted"), and are considered to be acutely undernourished. Wasting represents the failure to receive adequate nutrition in the period immediately preceding the survey and may be caused by recent illnesses or infections, causing loss of weight and the onset of undernutrition. Wasting may also reflect acute food shortages, which is accompanied by the children's low food intake . Episodes of wasting can also be seasonal where there is variation in the food supply or disease prevalence. With the correct intervention programmes wasting, which develops very rapidly, can also be reversed rapidly (WHO Working Group (1986)). Children whose weight-for-height index is below -3SD from the median of the reference population are considered to be "severely wasted".

The **weight-for-age** index is a composite of both the weight-for-height and the height-for-age indices, that takes into consideration both acute and chronic undernutrition. A weight-for-age Z-score of less than -2SD is classified as underweight.

Waterlow et al. (1977) suggest that weight-for-height be used as an indication of present nutrition, and height-for-age as an indication of past nutrition. When considering weight-for-age, they point out "that although weight-for-age has for many years been a mainstay in the evaluation of nutritional status, it has the disadvantage that it does not distinguish between acute and chronic malnutrition." The World Health Organization (WHO Working Group (1986)) endorse the use of a combination of the height-for-age and the weight-for-height indices -"In most circumstances separate indices should be constructed of weight-for-height and height-for-age, in addition to or in place of the classical index, weight-for-age.". The use of the weight-for-age index is advocated when length measurements are not performed accurately, which could occur as a result of the difficulty of measuring children under 5 years. "In some cases the choice might be highly influenced by technical errors in measurement, which affect some indicators more than others (e.g. weight-for-height is relatively unaffected by age errors but incorporates errors in the measurement of weight and height together) Pelletier(1991).

One of the more surprising results from the anthropometric indicators is that height-for-age and weight-for-height often do not show any association. Anderson(1979) found no correlation between the two indices in the results from five countries for children under 5 years. This indicates that stunted children are as likely to be wasted as non-wasted. Biologically one can gain height but one cannot lose it, whereas one can gain and lose weight; a child should treble its weight in the first year, but only double its height; catchup in height takes much longer than that in weight. An explanation given by Haaga (1986) for this lack of association between height-for-age and weight-for-height is that there is a random measurement error in height. This would reduce the association but not eliminate it. An approximate analysis suggests that a random measurement error would induce a very slight negative correlation between height-for-age and weight-for-height. This is indeed observed in the NH&NS data, since a significant but very small correlation of -0.03941 between height-for-age and weight-for-height is obtained.

5.5 ERRORS ARISING DUE TO CUT-OFF POINTS

The WHO Working Group (1986) discuss the criticism levelled at the use of cut-off points to separate "malnourished" from "normal". It was felt that cut-off points should be based on biological considerations, such as functional impairment and risk of mortality. Even though risk of death is not the only outcome that is considered, "the quantitative relation between mortality risk and anthropometric deficit will vary, among other things with infectious load.". Thus, determining a cut-off point on the basis of its ability to predict functional impairment may not even be generalisable from region to region.

Graitcer and Gentry (1981) also stress this approach:- "An international reference is useful in describing, in epidemiological terms, the overall growth of a country's or region's preschool children. The decision of whether or not to classify various segments of that population as malnourished, wasted, or stunted must, however, be based on external factors such as availability and amounts of resources rather than arbitrary anthropometric cut-off points.". For example if there is an emergency situation where resources are not readily available, a lower cut-off point might be necessary to identify those sections of the population most in need. Therefore, for example, one would perhaps look at the percentage below a cut-off of -3 to identify the stunted section of the population instead of using the generally accepted cut-off of -2. Keller(1991).

Waterlow(1990) uses an example of Davies(1988) to explain his misgivings about the use of prevalence below a cut-off point to measure stunting. In the study Davies shows the effect on the prevalence of stunting by adopting an Asian reference (from well-to-do children in different Asian countries) and compares it with the prevalence of stunting using the NCHS reference population. The prevalence is reduced from 50 percent with the NCHS reference to 16 percent with the Asian reference. Davies attributes the differences in the reference distributions to ethnic and genetic differences. Waterlow(1990) explains these misgivings:- "I have no way of telling whether Davies' view of ethnic differences is correct. The real point is that prevalences below a cut-off point are unstable; they may be useful for advocacy, but have little scientific value as a method of expressing results." In conclusion Waterlow comments that "it is better to avoid 'prevalence of stunted children' and instead to look at **changes** in the mean and distribution of height-for-age Z- scores".

Waterlow (1977) recommends that in addition to using Z-scores to express prevalence in malnutrition, they also be used to express population distributions, as they have a statistical meaning. Pelletier(1991) discusses the advantages and disadvantages of using prevalence to express malnutrition, and encourages the use of means in certain circumstances. Prevalence below a Z-score cut-off point of -2 has advantages in that as a practical measure, it appeals to policy makers and planners, which will also direct the policy attention towards the segment of the population of greatest public health concern. In many circumstances it might be more important for the policy makers to know whether the size of the high-risk group is changing, instead of whether the mean for that population is changing. The disadvantages of prevalence when compared with means however, is that statistical tests have more power to identify differences in the means of distributions of the nutritional indicator than in the proportions below a cutoff. In the case of low prevalence, which is often the case with the weight-for-height indicator, statistical tests for prevalences are much less powerful than those using means.

5.6 MEASURING ERRORS IN ANTHROPOMETRY

There are various types of measurement errors that can arise in measuring weight and height in nutritional anthropometry. They arise mainly as a result of errors due to inadequate training of the examiner, instrument errors, and measurement difficulties (Johnston 1981). It has been shown that systematic errors in measuring weight and height have a much larger impact on results than random errors (Pelletier 1991). Various recommendations have been made to minimize these errors. :-

- Training of personnel to use standardized, validated techniques, and instruments that are precise and correctly calibrated (Lohman et al. (1988))
- The precision can be assessed by examiners repeating the measurements on the same subject and calculating the inter- and intra- examiner standard deviations.
- To improve precision, the measurements should be performed in triplicate and the mean reported. A measurement should be investigated if it is too far from the mean of the other two.
- The accuracy of the measurements can be compared with those made by the supervisor (Gibson 1990).

The following common errors in measuring weight, length, and height are identified by Zerfas(1979) with the proposed solutions. Length (cm) refers to recumbent length i.e.

measurements of linear growth when the child is lying down. Stature (cm) refers to measurements made with the child standing up. Height may refer to either recumbent length or stature measurements. Weight was calculated in grams.

Length

Common error		**Solution**
Incorrect method for age	-	Use only when subject is < 2 years old
Footwear or headgear not removed	-	Remove as local culture permits (or make allowances)
Head not in correct plane	-	Correct position of child before measuring
Child not straight along board and/or feet not parallel with movable board	-	Have assistant and child's parent present; don't take the measurement while the child is struggling; settle child

Stature

Common error		**Proposed solution**
Incorrect method for age	-	Use only when subject is >= 2 years old
Footwear or headgear not removed	-	Remove as local culture permits
Head not in correct plane, subject not straight, knees bent, or feet not flat on floor	-	Correct technique with practice and regular retraining. Provide adequate assistance. Calm non-cooperative children

Board not firmly against head	Move head board to compress hair

Weight

Common error	Proposed solution
Room cold, no privacy	Use appropriate clinic facilities
Scale not calibrated to zero	Re-calibrate after every subject
Subject wearing heavy clothing	Remove or make allowances for clothing
Subject moving or anxious as a result of prior incident	Wait until subject is calm or remove cause of anxiety (e.g. scale too high)

The above procedures were followed in the child nutrition section of the NH&NS. In Section 5.9 it can be seen that there are systematic errors (where an entire PSU's Z-scores are over/underestimated) for an entire PSU in some cases. This is due to the interviewer not following the above procedures correctly.

5.7 AGE REPORTING ERRORS

Often age reporting is problematic in developing countries and the examiner or interviewer has no way of collecting this information. Cameron(1991) had the following comments to make on this issue. "Children, particularly in rural African areas, simply do not know when they were born or indeed how old they are. Age, per se, is of no practical importance to such children or their families; it is not celebrated with the passing of the seasons and is usually only known if a bureaucracy has arisen out of religious or governmental need. Thus age-independent measurements assume major importance, and the search for such measurements is constantly in the mind of the nutritional or growth researcher dealing with this problem.".

There have been several techniques proposed to compensate for this problem. These techniques include using "special events calendars" that include, for example religious, climatic, or social occurrences.

The techniques used .o collect the age information also have a significant effect on the classification of nutritional status. Gorstein (1989) demonstrates how the nutritional status of a population becomes classified differently depending on the technique used to define ages. The three techniques compared are:

(1) **The calculation of absolute ages,** where the date of birth of the child is known and the age is calculated by using the date of the visit/interview. An individuals age is calculated by recording the birth date and visit date, and the actual age can be calculated afterwards using a computer or pencil and paper. This approach can prove to be complicated in the field.

(2) **The rounding off to the nearest month.** By rounding to the nearest month, the nutritional status of children whose actual ages are less than the month is underestimated, while the status of children who are older than the month are overestimated. For a population with an even distribution of ages, this effect can be balanced but it will increase the variance. In a study by Gorstein(1989) it was found that the nutritional status of individuals were found to be overestimated, and this therefore influenced the levels of malnutrition for that population. The NH&NS used this method to determine age.

(3) **Truncating to the most recently attained month.** This method which simplifies the process of rounding for the interviewer, has the disadvantage of overestimating the nutritional status of all individuals.

Gorstein(1989) warns that the " improvement or degradation in nutritional status of populations might in fact not have a biological basis, but may be the result of the use of different statistical methodologies." The issue is therefore not whether any of the above three methods of age reporting are preferable, but rather that when any one of the methods are used, it is important that the method is documented and taken into account when interpreting and presenting the data.

There were not many nonresponses for the age variable in the NH&NS, and no serious problems were reported in collecting this information during the survey. According to the survey manager, the random re-interviews performed by the various supervisors in the field indicated that the age data that was collected was reliable and accurate. Therefore the assumption is made in all further discussions and analysis of this data, that the age data is reliable and accurate.

5.8 "DATA CLEANING" PROCEDURES USED IN THE NH&NS

As is described above, many anthropometric errors can arise in the field if the fieldworkers or interviewers do not follow the correct procedures. Various adjustments to the NH&NS anthropometric data, using the NCHS/CDC reference population, were designed and tested, in order to clean the data prior to analysis.

Univariate "data cleaning"

Height data

(i) The height-for-age data was analyzed for Z-scores between -6 and +6 (Z-scores outside this range are considered invalid (Dibley et al. 1987)). It would be expected that those invalid entries outside this range are individual observations randomly distributed over all the PSU's. In this case the status flag was set to "b".

(ii) The next scenario investigated was, where for a given survey round, an entire PSU's Z-scores were less than -3 and greater than -6. This scenario is extremely unlikely since one would not expect there to be such extreme stunting for an entire PSU. All height measurements in this case were excluded due to a systematic error (an unreliable interviewer) and the flag variable 'status' was assigned the value "e", indicating that the observations were to be excluded.

The height data for PSU=03127 in rounds 2 and 3 also had to be excluded, indicating that the interviewer (which in most cases was the same person from round 2 to round 4 within a PSU), was biased in collecting the anthropometric height data over a period of time.

(iii) In the case where 40% or more observations from a PSU have Z-scores larger than 3, the status flag was also set to "e". In a developing country like Lesotho where stunting is prevalent, one would not expect to observe 40% of children being overnourished. It was also found that the PSU's that fell into this category are predominantly from rural areas where malnutrition is highly prevalent, and therefore such a situation is even more unlikely.

(iv) The next adjustment to the data was to calculate the interquartile ranges of each PSU in each round, and set the status flag to "i" for all of those PSU's with interquartile ranges larger than 5 Z-score units. This adjustment seems necessary since the PSU's in these categories have a large proportion of Z-scores in the extreme tail areas, which is not characteristic of a developing country. One would question the accuracy of such data and may need to exclude it from the analysis. An example of such a case is PSU 04109 in round 3 (see Table 5.1), where an interquartile range of 5.6 is observed. Here one can observe that 25% of the Z-scores are larger than 3 and 25% smaller than -3 (i.e. 50% of the observations lie in the extreme tail areas), which is not what one would expect.

Table 5.1 Frequency and percentage of height for age Z-scores for PSU=04109 round 3

psu=04109 round=3	Z < -3	-3≤Z<-2	-2≤Z<+2	+2≤Z<+3	Z ≥ 3
frequency	6	2	8	2	6
percent	25%	8.3%	33%	8.3%	25%

The following frequency table is observed for the variable status after the above adjustment procedures are applied to the data.

Table 5.2 Frequency table of the variable "status" representing the height-for-age index for all the observations

STATUS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
a	8568	86.2	8568	86.2
b	286	2.9	8832	89.1
e	586	5.9	9418	95.0
i	199	2.0	9617	97.0
z	301	3.0	9940	100.0

where the status variable denotes:

"a" for the data points remaining after all the adjustments;

"b" for the data points excluded for Z-scores larger than 6 or smaller than -6, and is the last adjustment which is made;

"e" for observations from PSUs in which an abnormally high percentage of observations were very high or low;

"i" where the interquartile range is larger than 5;

"z" for missing height observations.

From Table 5.2 it can be observed that 1372 or 13.8% of the original height observations are not accurate enough or available for analysis i.e. where status = b,e,i, or z. It is necessary to report this percentage when the analysis is performed on the usable data.

Weight data

The weight data did not require as many adjustments as the height data and it seems that the fieldworkers coped better with the collection of these data. One entire PSU was excluded, namely PSU 04526 from round 3. In addition to having over 40% of it's Z-scores larger than 3 (see Table 5.3), the interquartile range was larger than 5. The flag variable status2 is set to "e" in this case. Therefore 15 of the 227 observations flagged "e" are from PSU 04526. Altogether only 5.8% of the weight data was unusable i.e. where status2=b, e, or z (see Table 5.4 below), compared with the height data where 13.8% is unusable.

Table 5.3 Frequency and percentage of weight for age Z-scores for PSU=04526 round 3

psu=04526 round=3	Z < -3	-3≤Z<-2	-2≤Z<+2	+2≤Z<+3	Z ≥ 3
frequency	0	0	8	0	7
percent	0%	0%	53%	0%	47%

Table 5.4 Frequency table of the variable status2 of the weight-for-age index for all the observations

STATUS2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
a	9348	94.3	9348	94.3
b	125	1.3	9473	95.5
e	227	2.3	9700	97.8
z	218	2.2	9918	100.0

Here status2 is defined the same way as "status" in Table 5.2.

In order to analyze the weight-for-height data, it is necessary to work only with that data that was valid for the weight data as well as the height data i.e. where status=a **and** status2=a. For both status=a and status2=a, there are 8443 observations or 85.1% of the data remaining for analysis of the weight-for-height data.

It is clear from the above process that it is necessary to inspect the data carefully before including or excluding an observation. One cannot merely include observations if its Z-score lies within a predetermined range. It is necessary to check for systematic errors made by interviewers who might be inexperienced or careless when measuring a child's height or weight. As is described above, this can be achieved by calculating the interquartile range of an interviewer's measurements, and using the (-3;3) range (for entire PSU's with invalid data) and the (-6;6) range (for random measurements that are invalid). Using a flag variable instead of merely deleting an observation is a safer practice and it enables one to provide a more detailed report regarding the excluded observations.

The disadvantage of the above approach is when the rules are applied to a PSU with a small sample size. In this case, data may be flagged for exclusion even though the measurement process was accurate.

Multivariate "data cleaning"

If one were to consider the Z-scores of the three indices simultaneously,

$$If\ \underline{x}_i = \begin{pmatrix} Z_{h/a\,i} \\ Z_{w/a\,i} \\ Z_{w/h\,i} \end{pmatrix} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \quad i=1,\ldots,n \qquad (5.1)$$

Denote the mean vector $\overline{x}$ and covariance matrix S of the x_i by:

$$\overline{x} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \overline{x}_3 \end{pmatrix} \qquad S = \begin{pmatrix} S_{11} S_{12} S_{13} \\ S_{21} S_{22} S_{23} \\ S_{31} S_{32} S_{33} \end{pmatrix} \qquad (5.2)$$

To test whether a particular observation $\underline{x}^*$ is outlying compute the quantity:

$$\chi^2 = (x^* - \overline{x})' S^{-1} (x^* - \overline{x}) \qquad (5.3)$$

Under the null hypothesis that x^* is not an outlier and assuming that the $\underline{x}_i$ are multivariate normally distributed, χ^2 is proportional to an $F_{3,df}$ random variable, where df is the degrees of freedom for S. For large df χ^2 is approximated by a χ^2_3 distribution. A significantly larger value of χ^2 may therefore indicate that $\underline{x}^*$ is an outlier. There is a linear dependency between weight-for-age, weight-for-height, and height-for-age. Had these three indices been perfectly linearly dependent then the covariance matrix would have been singular and the test statistic in (5.3) would not have been appropriate.

By computing the probability value for χ^2_3 for each observation x^* in the data set the outlying observations can be detected. Using this technique, for probability values less than 0.05, approximately 900 of the original 9900 observations were excluded from the analysis (see Table 5.5).

This technique has the advantage over the univariate technique described above in that all the variables can be analyzed simultaneously. Including age in the univariate approach described above would have complicated matters, since it would have been difficult to determine whether the outlying observations were due to age, weight, or height. Since the assumption is made that the age data is accurate, the univariate approach also allows one to include weight data where the height data is excluded or visa versa. From Table 5.5 it can be seen that the multivariate approach results in 9000 observations being accepted as valid for the three indices height-for-age, weight-for-age, and weight-for-height. The SAS program to perform this multivariate data cleaning procedure can be found in Appendix I.

Table 5.5 A comparison of the number of observations found to be valid after applying the two data cleaning procedures on the original 9900 observations

	height-for-age	weight-for-age	weight-for-height
univariate	8568	9348	8443
multivariate	9000	9000	9000

The multivariate cleaning procedure is an explicit significant test related to a p-value of 0.05, whereas the univariate test is based on the knowledge of poor observer technique. From the discussion on errors, in Sections 5.6 and 5.7 above, it is clear that most arise from poor observer technique. Given that, usually, each PSU had only one observer, the cleaning strategy was designed to identify PSU's with suspicious looking data in terms of the distribution of values for each measure within the PSU. Once identified, **all** the observations from that observer had to be excluded, not just the outliers, since one has to assume that the data within the acceptable range are also inaccurate, given the observer's poor technique. If there were no systematic errors resulting from poor observer technique that necessitated a univariate data cleaning process, a multivariate analysis using an approximate chi-squared test could have been used. It is for this reason that the univariate data cleaning procedure was applied to the NH&NS data and used in all the analyses in Chapters 7, 8 and 9.

CHAPTER 6

THE ESTIMATION AND PORTABILITY OF STANDARD ERRORS AND DESIGN EFFECTS OF RATIO ESTIMATES USING DATA FROM A COMPLEX SURVEY DESIGN

6.1 INTRODUCTION

A requirement of most surveys is to calculate sampling errors for each estimate. The estimation procedure must take into account the design of the survey, in particular its clustering and stratification. At the same time, simple computational formulae should be provided so that an economically feasible procedure is established to produce the numerous estimates in a survey (Verma 1982).

The Yugoslavia Federal Statistical Office(1978) made the following remarks with regards to sampling errors:
"Data from a sample survey might be, at least in principle, anything between 'excellent' and 'useless'. An inspection of the magnitude of sampling errors for various characteristics at the level of the country as a whole as well as its subdivisions is the first step in passing judgement about the place of the survey between these extremes. Therefore in order to establish the basis for the evaluation process of the sample survey data, information about the magnitude of sampling errors should be considered as an indispensable part of each sample survey report."

Many household health surveys are still analyzed using the survey scheme that was developed by the Expanded Programme on Immunization (EPI) of the World Health Organization (WHO) (Henderson and Sundaresan 1982, Lemeshow and Robinson 1985). The EPI survey, often referred to as a "30x7" survey, is a cluster-sampling scheme in which 30 clusters each consisting of 7 children is chosen. This made it possible to estimate immunization coverage to within 10% of the true proportion. There have been certain misgivings about using the EPI strategy. Ferrinho et al. (1992) had the following comments to make in this regard - "Without a clear understanding of the limitations of the sampling strategy used, this sampling strategy has been extended to other types of surveys." These misgivings are due to the clustering effect which is implicit in the strategy which are difficult to extrapolate to variables other than immunization status. (Bennet et al. (1991), Ferrinho et al. (1992), and Lemeshow et al. (1985).

Therefore a more suitable and generalized technique for cluster sampling with the option of stratification, which originates from Kish(1965), is considered by Bennet et al. (1991). Standard statistical packages such as SPSS X (1995) and SAS (1994) do not cater for complex survey data. Therefore it might not be surprising to discover that many surveys with complex survey designs have been analyzed without adjusting for clustering or stratification. This would have resulted in incorrect and smaller standard error estimates and misleading results being reported.

Kish et al(1976) introduced the concept of 'portability', which 'refers to the possibility of carrying over from one subclass to another, from one variable to another or from one survey to another, the conclusions drawn regarding the sampling error' (Verma 1982). Portability can be applied to the standard error, the design effect, and the rate of homogeneity (ROH) (Verma 1982). The different measures are portable to different degrees. Portability is recommended for a survey such as the NH&NS when the results of the survey are generated. This procedure could have the effect of saving much time and money if applied correctly.

In this chapter procedures are provided for calculating standard errors, under a complex sampling design such as the NH&NS, and also design effects, under the assumption that the response variables are dichotomous in nature (Ratio estimates are commonly encountered estimates from sample surveys, and means and proportions are just special cases of ratio estimates (see Section 6.2)). A user-friendly program is written in SAS to produce estimates of the standard error and design effects, as well as a 95% confidence interval for the ratio estimator. This program, called STANDERR is presented in Appendix A. The process of portability is then assessed and applied to the NH&NS.

6.2 THE ESTIMATION OF THE STANDARD ERROR OF A RATIO ESTIMATOR

It should be noted that **without replacement sampling** (wor) was used in the first sampling stage of the NH&NS since it would have been undesirable if any PSU would have been chosen more than once. However variance estimation in this chapter is based on **with replacement sampling** (wr) since under wor sampling the formula for the variance becomes very complicated requiring cumbersome computer programs (Stoker et al. 1997). This approach is based on the reasoning of Kish(1965) that wor sampling be used in all sampling stages even though the estimation of the variance is applicable to wr sampling. Although this approach tends to overestimate the true variance of an estimator, the overestimation is small in general. Furthermore, Stoker et al. (1997) compared the variance estimates under wor and wr and concluded that using the variance estimate based on wr sampling is in general a conservative and safe procedure under wor sampling.

For many surveys the results are presented in the form of proportions. A proportion is the mean of a dichotomous variable, where members belonging to a specific class receive the value $Y_i = 1$, and $Y_i = 0$ for non-members. This is referred to as a binomial variable. For multiple classification, it is referred to as a multinomial variable. Therefore p=y/n is the proportion of sample elements that belong to the defined class, and q=1-p is the proportion that do not belong.

For simple random sampling (srs) the usual formula for the variance is

$$var(p) = (pq/n) \qquad (6.2.1)$$

Consider a sample design with L strata in which m_h PSUs are drawn from the h-th stratum. Let y_{hi} and x_{hi} respectively denote the estimated total values of the variables or characteristics y and x for the i-th PSU drawn from the h-th stratum, obtained from the sample data. The estimated total values y_{hi} and x_{hi} of y and x are:

$$y_{hi} = \sum_j w_{hij} y_{hij} \quad , \quad x_{hi} = \sum_j w_{hij} x_{hij} \qquad (6.2.2)$$

where y_{hij} and x_{hij} respectively denote the values of y and x for the j-th sample element belonging to the i-th PSU drawn from the h-th stratum, with w_{hij} the weight associated with this sample element.

The stratified ratio estimator using the above notation is defined as

$$r = \frac{\sum_{h=1}^{L} \sum_{i=1}^{m_h} y_{hi}}{\sum_{h=1}^{L} \sum_{i=1}^{m_h} x_{hi}} \qquad (6.2.3)$$

where L is the number of strata and m_h the number of PSU's drawn from the hth stratum. Averages and proportions are special cases of ratios, where in both cases x is an indicator or counting variable with

x_{hij} = 1 if the j-th sample element drawn from the i-th PSU drawn from the h-th stratum must be included in the calculation.

x_{hij} = 0 if this is not the case.

y_{hij} = 1 if the j-th sample element drawn from the i-th PSU drawn from the h-th stratum possesses a particular characteristic.

y_{hij} = 0 if this is not the case or if x_{hij} = 0.

Since the ratio estimator (6.2.3) is summed over the entire sample, we could use (6.2.1) and calculate the variance as if it were from a simple random sample so that

$$var(r_{srs}) = r(1-r)/n \qquad (6.2.4)$$

In order to estimate the variance of a ratio (proportion) for a multistage sampling design when it includes both stratification and clustering, consider the following: Let S_r^2 denote the variance of the stratified ratio estimator r, then (Kish (1965)):

$$\begin{aligned} S^2{}_r &= \frac{1}{x^2}\Big(\sum_{h=1}^{L} var(y_h) + r^2\sum_{h=1}^{L} var(x_h) - 2r\sum_{h=1}^{L} cov(y_h, x_h)\Big) \\ &= \frac{1}{x^2}\Big(\sum_{h=1}^{L} d^2y_h + r^2\sum_{h=1}^{L} d^2x_h - 2r\sum_{h=1}^{L} dy_h dx_h\Big) \qquad (6.2.5) \end{aligned}$$

where

$$x = \sum_{h=1}^{L} x_h = \sum_{h=1}^{L} \sum_{i=1}^{m_h} x_{hi}, \qquad (6.2.6)$$

$$y_h = \sum_{i=1}^{m_h} y_{hi}$$

Kish(1965) uses the symbol d to denote concisely the variance terms of the sample sums within strata, each based on m_h random selections, which can be computed as

$$d^2 y_h = \frac{(1 - f_h)}{m_h - 1} (m_h \sum_{i}^{m_h} y^2{}_{hi} - y_h{}^2),$$

$$d^2 x_h = \frac{(1 - f_h)}{m_h - 1} (m_h \sum_{i}^{m_h} x^2{}_{hi} - x_h{}^2),$$

$$dy_h dx_h = \frac{(1 - f_h)}{m_h - 1} (m_h \sum_{i}^{m_h} y_{h_i} x_{h_i} - y_h x_h)$$

(6.2.7)

Using the notation:

$$z_{hi} = y_{hi} - r x_{hi} \qquad (6.2.8)$$

and

$$z_h = \sum_{i=1}^{m_h} z_{hi} \qquad (6.2.9)$$

it can easily be shown that (6.2.5) can be written:

$$S^2{}_r = \frac{1}{x^2}\sum_h^L d^2 z_h, \textit{ where}$$

$$d^2 z_h = \frac{(1 - f_h)}{m_h - 1}\,(m_h \sum_i^{m_h} z^2{}_{hi} - z_h{}^2) \qquad (6.2.10)$$

so that S_r^2 and the standard error S_r is estimated as

$$S_r^2 = \frac{\sum\limits_{h=1}^{L} \frac{1 - f_h}{m_h - 1}(m_h \sum\limits_{i=1}^{m_h} z_{hi}^2 - z_h^2)}{(\sum\limits_{h=1}^{L} x_h)^2},$$

$$S_r = \sqrt{S_r^2}$$

(6.2.11)

6.3 CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

The 95% confidence interval for a population proportion is given approximately by

$$(r - 1.96S_r \;;\; r + 1.96S_r) \qquad (6.3.1)$$

provided that the sample is large i.e. the number of PSU's which are drawn is large, in which case r is approximately normally distributed (Stoker 1988). Ratio estimates can be biased. In order to keep this bias small the coefficient of relative variation of the variable in the denominator is calculated. If this coefficient is less than 0.10 the bias is ignored in practice. If not, the estimated results could be biased and misleading.

6.4 THE DESIGN EFFECT

The design effect (deff) is defined as the ratio of the variance of the estimator, when the complexity of the sample design is taken into account, to the variance of the estimator when it is assumed that the sample data were obtained by means of a simple random sample. The combined effect of stratification and clustering on the variance of an estimator of a population characteristic is measured by the design effect. From (6.2.4) and (6.2.11)

$$deff = S_r^2/var(r_{srs}) \qquad (6.4.1)$$

If deff is greater than one, the variance of the estimator of the population characteristic is larger than under simple random sampling in which case complex sampling is less efficient than simple random sampling. If deff is less than 1, complex sampling *is more efficient than random sampling. Put in another way, if* by mistake or due to lack of experience and knowledge (which is common in a developing country), the srs estimate of the variance ($var(r_{srs})$) is used instead of the variance (S_r^2) for a complex survey design, then deff is the measure of the underestimation or overestimation for using the incorrect variance.

The value of the design effect (deff) tends to increase with increasing cluster size for a given variable, a given number and type of cluster, and subsampling procedure used. Kish (1965) introduced a synthetic measure roh (rate of homogeneity) to control this effect, which is defined in terms of deff as

$$deff = S_r^2/var(r_{srs}) = 1 + roh(\bar{b} - 1) \qquad (6.4.2)$$

where $\bar{b}$ is the average cluster size.

For a single stage cluster sampling design, **roh** is the measure of homogeneity within clusters which tends to increase with the variance of the sample. It is also known as the coefficient of intraclass correlation (Kish 1965). For multistage complex surveys, roh is composed of the variability of all the stages of the survey design. It can also be regarded as measuring the effect of clustering within strata for such surveys. It has been shown that larger clusters tend to have smaller values for roh (Cochran 1977) since "neighbouring units tend to resemble each other. Then it follows that the resemblance decreases with increasing distance over the distribution" (Kish 1965). Bennet(1991) gives various guidelines for estimating roh, based on similar types of questions from past surveys. The value of roh is usually less than 0.4. The comment is made however that these estimates of roh are necessarily vague "as there will be variability in the value of roh from country to country, from survey to survey, and from item to item". Other reasons given for varying sizes of roh are the varying levels of efficiency of the interviewers and the supervisors. The recommendation is made that, if possible, roh should be chosen from an earlier round of the same survey (see portability of roh in Section 6.6).

6.5 THE COMPUTER PROGRAM - STANDERR

Nearly all surveys in developing countries are stratified clustered samples, with clusters of households in each of the regions or strata. These survey designs reduce the cost of data collection, but in exchange complicate the data analyses, since the observations are no longer independent and identically distributed (iid). It would be interesting to know (and difficult to find out) how many researchers have produced misleading results in the past by underestimating the variance of the estimates due to not taking the complexity of the survey design into consideration. Standard statistical packages such as SAS (1994) and SPSS (1995) assume iid observations drawn from infinite populations and hence are not appropriate for most survey designs. Bennet et al. (1991) make the following remarks concerning complex survey designs in developing countries: "There are many excellent textbooks which describe complex designs and appropriate formulae for their analysis, but a certain level of expertise is needed to make the most of these, and this is often not available to workers in the field.". One way to develop this expertise is to have a user friendly computer program at one's disposal to handle such data. Computer programs which have been developed for complex survey designs are :

1) PSALMS - a program which is available only under the OSIRIS IV program package of the University of Michigan. This package is written for a mainframe computer and is therefore expensive, and generally inaccessible to most developing countries.

2) CLUSTERS - a program which was developed for the World Fertility Surveys - see Verma et al. (1978). This program is more freely available, since there is no copyright on its use. However, it is more inconvenient to use than PSALMS (see Stoker(1988)).

3) A recent computer package to become available for handling complex data is STATA (STATA Corporation (1993)). This program facilitates regression type analyses under clustering, but it does not make provision for stratification (see Chapter 5). There are also no procedures to handle log-linear modelling from a complex survey design in this package.

4) Another program called SUDAAN (Sudaan, 1995) is available which does facilitate clustering and stratification. It has procedures for handling log-linear modelling, logistic regression, linear regression, and various other statistical techniques. It has been used in this thesis for testing independence in two-way tables and for fitting a logistic regression model. SUDAAN does however have certain limitations when analysing categorical data from a complex survey design, which are discussed in Chapter 7 (Section 7.6)

A program called STANDERR has been written in SAS by the author for a complex survey design (see Appendix A). SAS is a well known and commonly used statistical package. Instead of a BOS having to convert data sets from the format of one computer program to another (e.g. from SAS to SUDAAN), STANDERR can be used. It must be remembered that problems arose in the NH&NS when data was converted from the DBASE data format to the SAS format (see Chapter 3). STANDERR calculates proportions, the design effect, the standard error, and 95% confidence intervals for any number of cells, for any number of dimensions from a contingency table. It only requires the user to specify the data set, the stratum variable, the PSU (cluster) variable, and the dimensions of each variable used in the analysis.

Consider the Lesotho NH&NS where the proportion of stunted children in the population under 5 years, by sex, is investigated. Table 6.1 is obtained by using the program STANDERR. The design effect (using equation (6.4.1)) for stunted females for example, indicates that the standard error estimated as if the sample were from a simple random sample, would be underestimated by a factor of nearly 1.414, compared to the standard error which takes the complex survey design into consideration. Ferrinho et al(1992) illustrate this point clearly using three case studies from Alexandra, South Africa. The design effect for certain variables relating to housing varied from 2 to as much as 6.99, indicating by how much the variance would have been underestimated if the complex survey design had not been considered when analysing the data.

In Chapter 7 the design effect is used as a correction factor to the chi-square test statistic for assessing the association between certain variables in two-way and multiway tables. It will be shown how the results are affected if the design effect is not considered in the analyses.

Table 6.1 Percentages, standard errors (for a complex survey design), and design effects of the stunted children by sex, in the Lesotho National Household Health and Nutrition Survey.

total percent standard error design effect	female	male
stunted	11.11% 0.49% 1.9244	12.82% 0.52% 2.3139
not stunted	38.02% 0.71% 1.8146	38.06% 0.68% 1.7526

6.6 PORTABILITY

In a large-scale multi-purpose survey, where estimates from cross-tabulations involving many cells are calculated it is a time-consuming task to calculate sampling errors for each estimate in each cell. This often leads to the slow processing and release of the results. In the NH&NS, for example, a cross-tabulation of education (8 categories) by district (10 categories) requires the calculation of 80 sampling errors for the 80 cells. In practice only an approximate value for any standard error need be calculated. This can be achieved by providing some means of extrapolation of errors from computations for selected variables and sample categories, to other variables and categories for which actual computation was not performed (Verma(1982)). In a survey such as the NH&NS, where the same survey design was repeated, the standard error or some statistics derived from it may be relatively stable from one survey round to the next. Thus the variance pattern in the June round could have been used to predict sampling errors for the subsequent rounds in November and February.

It might also not be feasible to publish all the sampling errors even if they were all computed. The number of tables could become too large. Therefore if the user is provided with a simple way of computing approximate sampling errors for any estimates in which he/she is interested, it will not be necessary to publish all the sampling errors. This again implies extrapolation, and a knowledge of the patterns of variation of sampling errors is required.

A knowledge of the patterns of variation of sampling errors can also be used to evaluate how a particular design has fared, which could assist in designing future samples (Verma (1982)).

Kish et al (1976) introduced the concept of portability to address these issues. Portability refers to the possibility of carrying over the conclusions drawn regarding the sampling error from one subclass to another, from one variable to another, or from one survey to another. Following this procedure correctly and cautiously could serve as another way to simplify the survey process and save a developing country with a tight budget much time and money in processing the results.

There are three types of subclasses : cross-classes, mixed classes, and geographic classes. **Cross-classes** are groups that are defined in terms of demographic characteristics e.g. age and sex tend to be uniformly distributed geographically across the population. **Geographic classes** are completely segregated into separate clusters, which implies that a whole cluster either belongs or does not belong to the subclass. **Mixed classes** are classes that are less well distributed than cross-classes and not as completely segregated as geographic classes. Particular ethnic, occupation, or socio-economic groupings can be considered mixed classes. For example from the NH&NS, the higher educated class of the population mainly fall in the urban areas although not exclusively so.

6.6.1 ASSESSING THE PORTABILITY OF THE STANDARD ERROR, DESIGN EFFECT AND RATE OF HOMOGENEITY

The standard error, the design effect and the rate of homogeneity (roh) are considered in the portability process. These measures are portable to different degrees.

A standard error (see (6.2.11)) depends upon a number of factors (Verma 1982):

1. the nature of the estimate;
2. its units of measurement (scale) and magnitude;
3. its variability in the population;
4. the sample size;
5. the sample design (clustering, stratification, weighting, cluster size, etc.);
6. the nature and size of the sampling units;
7. for sample subclasses, their nature and distribution across sample clusters.

In order to enhance the portability of the standard error across subclasses, variables and sample designs, the effect of the above factors have to be reduced (Verma 1982). This process will be described in Section 6.6.2.

DESIGN EFFECT

The design effect (deff) as defined in Section 6.4 as

$$deff = S_r^2/var(r_{srs})$$

is more portable than the standard error since it does not depend upon factors which affect both ($var(r_{srs})$) and (S_r^2) in the same way. These factors include the units of measurement, magnitude of the estimate, its variability in the population, and sample size. Deff depends upon factors such as the nature of the estimate, the sample design and the type and size of the sampling unit (Verma 1982).

RATE OF HOMOGENEITY (ROH)

For a given variable and a given number and type of clusters and subsampling procedure used, the value of the design effect tends to increase with increasing cluster size (Verma 1982).

The rate of homogeneity defined in Section 6.4 as

$$deff = S_r^2/var(r_{srs}) = 1 + roh(\bar{b} - 1)$$

which measures the degree of correlation between members of a cluster, removes the effect of the average cluster size in deff (see Verma 1982).

The basic assumption is made by Verma(1982) that roh depends upon the nature of the variable, so that the relative values of roh for two variables persist as one moves from the total sample to diverse subclasses and possibly to other sample designs. Previous studies by Verma et al. (1980) supported this assumption where it was found that the ranking of median rohs for groups of substantively similar variables were consistent across a number of World Fertility Surveys.

6.6.2 THE INFERENTIAL PATH FOR THE PORTABILITY PROCESS

Since the sample size of a subclass is smaller than the total sample size, and since the standard error depends upon the sample size, one cannot directly impute a standard error from the total sample to a subclass. The general inferential path in this case, is from the standard error (se_t) computed for the whole sample to the design effect of the whole sample ($deff_t$) to $deff_s$ and finally to se_s the standard error for the subclass (see Section 6.6.2).

The subclass design effect ($deff_s$) is expected to be smaller than that of the total sample since there is a smaller effect of clustering in the subclass. Therefore in order to relate the design effect of the total sample to the design effect, and since the relative values of roh for two variables persist as one moves from the total sample to diverse subclasses, a model based on equation (6.4.2) is

$$\frac{(deff_s - 1)}{(deff_t - 1)} = \frac{roh_s\ (\bar{b}_s - 1)}{roh_t\ (\bar{b}_t - 1)} \qquad (6.6.1)$$

where the subscript s indicates a subclass, t indicates the total sample, $\bar{b}_s$ is the average cluster size for the subclass, and $\bar{b}_t$ is the same for the total sample.

For **cross-classes**

$deff_s = 1 + M_s(deff_t - 1)$ from (6.6.1) (6.6.2)

where $M_s = (\bar{b}_s - 1)/(\bar{b}_t - 1)$ is the average size of the subclass relative to that of the whole sample.

In (6.6.2) it is assumed that $roh_s = roh_t$ because the nature of the sampling units and the sampling procedure have not changed and therefore roh is portable between the total sample and the subclass (Verma 1982).

For **geographic classes**

$deff_s = 1 + c_s(deff_t - 1)$ from (6.6.1) (6.6.3)

where $c_s = roh_s\ (\bar{b}_s - 1)\ /\ roh_t\ (\bar{b}_t - 1)$

is a constant for the domain to be determined empirically by fitting the above relation to the computed $deff_s$ over a group of variables.

For **mixed classes**

$deff_s = 1 + M_s^{\alpha}(deff_t - 1)$ from (6.6.1) (6.6.4)

where M_s has the same definition as for (6.6.2) and α is an empirically determined parameter expected to lie in the range 0 to 1, with values at the upper end corresponding to cross-classes, and at the lower end to segregated or geographic classes (Verma (1982)).

The inferential path is from standard error (se_t) for the whole sample to the corresponding $deff_t$, to $deff_s$ for the subclass and finally to the standard error (se_s) of the subclass.

In order to relate the deff's to the standard errors, the following expressions are used

$se^2_t = (s^2_t / n) deff_t$ for the total sample (6.6.5)

and $se^2_s = (s^2_s / n) deff_s$ for the subclass (6.6.6)

where s^2_t and s^2_s are equivalent to $var(r_{srs})$ in (6.4.1) and se^2_t and se^2_s are equivalent to S_r^2 in (6.4.1).

The relationships (6.6.5) and (6.6.6) hold when the sample is not self-weighting. For very small subclasses the effect of clustering and stratification tends to disappear, i.e. the design effect tends to 1.0. However the effect of sample weighting tends to persist, and in order to compensate for the unequal weights, Kish (1965) introduces a loss factor

$$L = \frac{\sum n_h w^2{}_h . \sum n_h}{(\sum n_h . w_h)^2} \qquad (6.6.7)$$

which is multiplied with the variance of all the estimates, where n_h is the number of units with weight w_h.

It is found in practice (Verma 1982) that deff for very small cross-classes tends to the value L in accordance with (6.6.7). A study by Verma et al. (1980) confirms this on a basis of a very large number of computations. Thus an adjusted design effect deff′ excluding the loss function is defined as

$$\text{deff} = (S_r^2)/(\text{var}(r_{srs})) = L \,.\, \text{deff}'$$

Therefore the standard errors for the total sample (se_t) and the subclass (se_s) can be written as

$$se^2{}_t = (s^2{}_t \;/\; n)\; L_t\; \text{deff}_t' \qquad \text{for the total sample} \qquad (6.6.8)$$

and $$se^2{}_s = (s^2{}_s \;/\; n)\; L_s\; \text{deff}_s' \qquad \text{for the subclass} \qquad (6.6.9)$$

Based on the relationships described above, the portability process is schematically represented in Table 6.2, where the direction of the standard errors, design effects and roh values from their computed to imputed values are shown.

Table 6.2 Schematic representation of portability

$$\begin{tikzcd}
\text{Computed SE} \arrow[d] & \text{Imputed SE} \\
\text{Computed deff} \arrow[r, "\text{subclasses}"] \arrow[d] & \text{Imputed deff} \arrow[u] \\
\text{Computed roh} \arrow[r, "\text{variables}"] & \text{Imputed roh} \arrow[u]
\end{tikzcd}$$

6.6.3 PORTABILITY APPLIED TO THE NH&NS

Portability was not used in the case of the NH&NS and therefore for the purposes of this thesis it will be shown how the sampling errors from the June round of the NH&NS could have been used in the November round. Quite complicated estimation procedures are proposed (Verma 1982) in the case of geographic and mixed subclasses. If not applied correctly, the portability process could lead to additional errors in the survey results.

Therefore portability might not have the desired effect of simplifying the survey process and saving a statistician in a developing country time in releasing the results of their survey. Based on these arguments portability is only recommended for cross-classes and repeated surveys in developing countries.

PORTABILITY ACROSS GENDER

An example of a typical cross-class is gender. In order to estimate the standard error of malnourished female children from the total population of children under 5 years consider the relationship in (6.6.2).

$$\text{deff}_f = 1 + M_f(\text{deff}_t - 1) \qquad (6.6.10)$$

where the subscript f indicates females, the subscript t indicates thr total sample, and where

$$M_f = (\bar{b}_f - 1)/(\bar{b}_t - 1)$$

Considering firstly females, it is assumed that $\rho_f = \rho_t$ because the nature of the sampling units and the sampling procedure have not changed from the total sample to the subclass of females. It is also reasonable to assume $L_f = L_t$ since the relative allocation among domains (n_h in equation 6.6.7) for the subclass of females is similar to that for the total sample.

From the data $\text{deff}_t = 2.74912$, $\bar{b}_t = 89.6$, $\bar{b}_f = 45.225$ and $s^2{}_f/n = 0.000022005$

Therefore from 6.6.10 $\text{deff}_f = 1 + (45.225-1/89.6-1)(2.74912-1)$

$$= 1.8736$$

and from (6.6.6) $(se_f)^2 = 0.000041228$

so that $se_f = 0.0064209$ which is reasonably close to the computed value of 0.0058551.

For males $\bar{b}_m = 44.36$, $s^2{}_m/n = 0.000024991$ and

$$\text{deff}_m = 1 + (44.36\text{-}1/89.6\text{-}1)(2.74912 - 1)$$
$$= 1.856$$

and from (6.6.6) $se_m = 0.0068105$ which is very close to the computed value of 0.0068487

PORTABILITY ACROSS SURVEY ROUNDS

In the NH&NS the different rounds of the survey had the same sampling designs, with slightly different average cluster sizes. If the portability process is applied in this case, one would not be able to simply "transport" the design effects from one round to the next since the average cluster size has an effect on the design effect and therefore roh should be used, since it removes this effect. We therefore assume that the value of roh (ρ) in the two rounds is the same.

If for example the standard error of the proportion of malnourished female children from the round 2 survey is to be used in the round 3 survey, the following procedure is applied:

Since there is a significant difference between the proportion of malnourished children in the rural areas and the urban areas (see Section 7.7), a separate portability process is necessary for each of these areas.

From (6.6.1) s=rur3 and t=rur2, where rur2 and rur3 indicate respectively the 2nd and 3rd survey rounds for the rural areas

$$\text{deff}_{rur3} = 1 + M_s(\text{deff}_{rur2} - 1) \qquad (6.6.12)$$

where $M_s = (\bar{b}_{rur3} - 1)/(\bar{b}_{rur2} - 1)$

since it is assumed that $\rho_{rur3} = \rho_{rur2}$

From the data $\text{deff}_{rur2} = 2.00224$, $\bar{b}_{rur3} = 24$, $\bar{b}_{rur2} = 23$,
and $s^2_{rur3} = 0.000010993$

Therefore $\text{deff}_{rur3} = 1 + (23/22)(2.00224 - 1)$

$$= 2.04796$$

and from (6.6.6) $(se_{rur3})^2 = 0.0000225$

so that $se_{rur3} = 0.004745$

A similar process is applied to the malnourished female children in the urban population, where $(se_{urb3})^2 = 0.0000535$ so that the combined variance of rur3 and urb3 is

$$(se_3)^2 = (se_{rur3})^2 + (se_{urb3})^2 + 2\text{cov}(rur3,urb3)$$

where cov(rur3,urb3) is the covariance between the malnourished female children for the rural and urban areas in the third survey round.

Since the rural area and the urban area data can be considered to be two independent samples (urban and rural areas are in different strata), cov(rur3,urb3)=0, and therefore

$$(se_3)^2 = (se_{rur3})^2 + (se_{urb3})^2$$

$$= 0.000076$$

and $(se_3) = 0.008718$

for the rural and urban areas combined.

Using the program STANDERR on the actual data in the third survey round, $se_3 = 0.008998$, which is not that different from 0.008718, and therefore the portability process is seen to produce a suitable estimate for the standard error in this case.

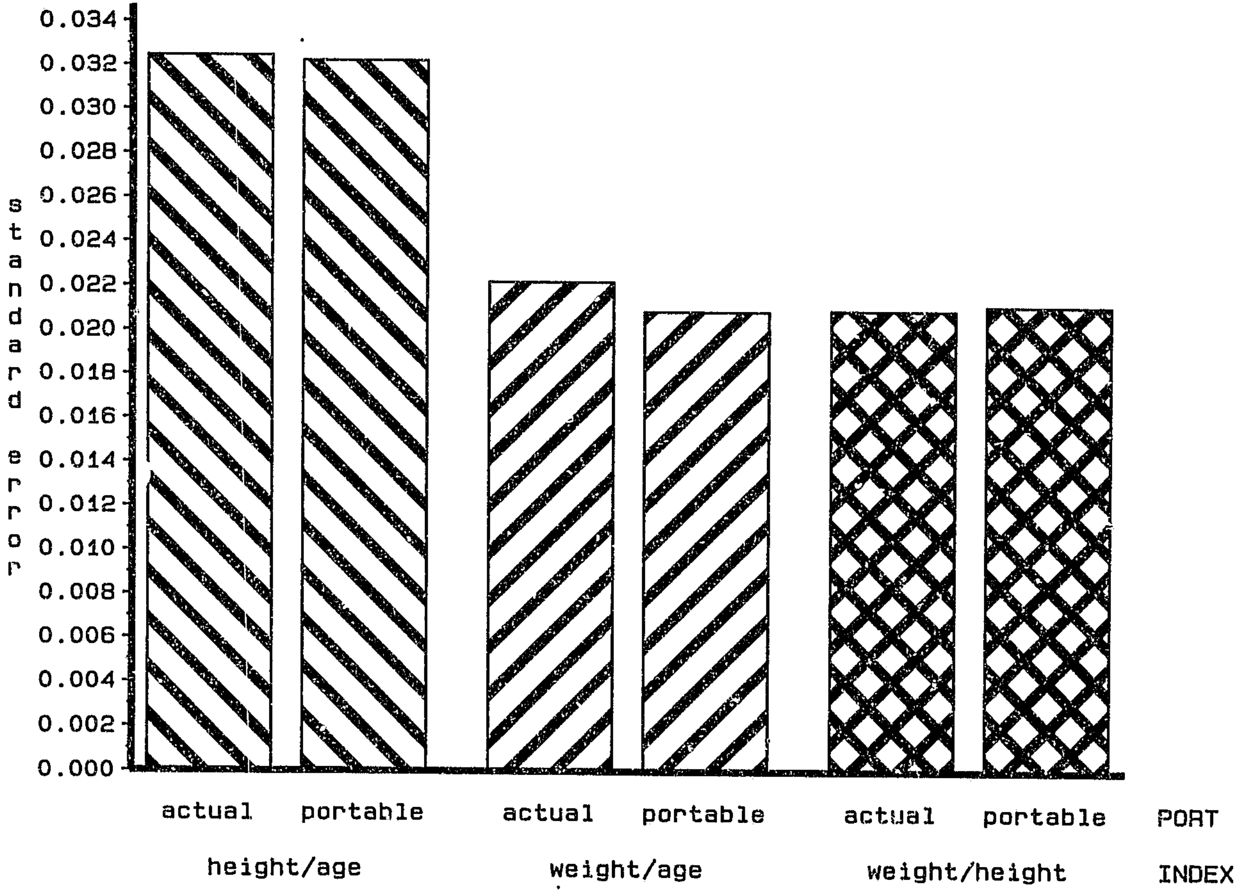
The same process would be applied to the remaining cells of the contingency table i.e. for malnourished males, adequately nourished females, and adequately nourished males. The above procedure might not be the best option to follow when there are just four cells in the table, as it might be quicker to estimate the actual standard errors. It is more appropriate when dealing with larger numbers of cells. In Figure 6.1 the effect of portability on all three of the indices discussed in Chapter 5 (height-for-age, weight-for-age and weight-for-height) across survey rounds is displayed. It appears that the 'actual' standard errors (using STANDERR) and the standard errors estimated by portability are all very simiiar. Thus, by applying portability, there would have been little effect on the magnitude of the standard errors (for all three indices) in the October survey, had the standard errors from the June survey round been used to estimate them.

6.7 SUMMARY

It is clear from the above examples that portability can save a statistician much time in having to produce standard errors and design effects when estimation takes place across subclasses, variables or survey rounds. It is not feasible to compute and present standard errors individually for proportions or means of all variables and subclasses in large surveys like the NH&NS. This makes portability an attractive option. In a multi-round survey such as the NH&NS, the estimated standard error of variables such as weight-for-height, weight-for-age and height-for-age, from the first round survey, can be used to estimate the sample size for the second round survey.

It should be noted though that the procedure should be practised with caution and logic. For instance it was necessary to apply portability to the rural and urban areas separately before combining them, since there is a significant difference in the proportions of malnutrition between the two areas (see Section 7.7).

FIGURE 6.1 Portability of the standard error from the June to October Survey round for the indices weight-for-age, weight-for-height and height-for-age



The computer program STANDERR described in this chapter will now enable a survey statistician analysing data from a complex survey design to:

- use the well known computer package SAS
- correctly calculate standard errors and design effects
- and analyze two-way or multiway tables.

The computer program STANDERR (Section 6.5) to estimate standard errors, confidence intervals and design effects from a complex survey design is used in most of the analyses in the remaining chapters of this thesis.

CHAPTER 7
THE ANALYSIS OF TWO-WAY AND MULTIWAY TABLES

7.1 INTRODUCTION

Normally it is possible to classify the members of a population in many different ways. People, for instance, may be classified into male and female; into married or single; into malnourished or not malnourished (as in the case of the NH&NS); and so on. These are examples of dichotomous classifications. Multiple classifications also are common when people are classified into left-handed, ambidextrous or right-handed; or blue-eyed, brown-eyed, green-eyed and others. Whether the classification is dichotomous or multiple it must be exhaustive, and the categories into which it divides the members of the population must be mutually exclusive. A classification is exhaustive when it provides sufficient categories to accommodate all members of the population. The categories are mutually exclusive when they are so defined that each member of the population can be correctly allotted to one, and only one category. When the observations from a sample have been classified in two separate ways, the results can be arranged in a rectangular table. For instance from the NH&NS (see Table 7.1) there are 3727 children under two years checked for stunting and breastfeeding. Of the 3354 children that are not malnourished, 2763 are still breastfed and of the 373 that are malnourished, 289 are still breastfed. A table such as this is known as a contingency table and the data is referred to as categorical data. This 2 x 2 table (i.e. 2 categories for stunting and 2 categories for breastfeeding) for example is the simplest form of a two-way table. Had the classifications been multiple rather than dichotomous the table, while still being square or rectangular, would have had many more cells.

Table 7.1 An example of a dichotomous (2 x 2) two-way contingency table from the NH&NS

	Malnourished=No	Malnourished=Yes	Total
Still Breastfed	2763	289	3052
Not Breastfed	591	84	675
Total	3354	373	3727

If the contingency table of Table 7.1 is further cross-classified by whether the mother of the child is educated or not, then there are three different classifications and the contingency table is known as a three-way table. For example from Table 7.2 there are 1907 children that are still breastfed, who are not malnourished and who have uneducated mothers. Higher than two-way tables are also often referred to as multiway tables.

Table 7.2 A three-way table from the NH&NS of education levels of mothers by breastfeeding and by nutritional status of children under two years

	Malnourished=No		Malnourished=Yes		
	Uneducated	Educated	Uneducated	Educated	Total
Still Breastfed	1907	856	225	64	3052
Not Breastfed	379	212	58	26	675
	2286	1068	283	90	3727

The three main techniques that are applied in this thesis to analyze data from two-way and multiway tables from the NH&NS are: the Wald chi-square test using the full covariance matrix under the sample design, correction factors applied to the ordinary Pearson chi-square statistic and log-linear modelling in the case of a multiway table (see Section 7.2). It is shown how these techniques compensate for a complex survey design such as the NH&NS. The programs used for analysis are STANDERR described in Chapter 6 and SUDAAN(1995).

Work by Rao and Scott(1984) and others have shown that clustering in a survey design can have a substantial impact on the significance levels of the standard Pearson chi-square statistic

(χ^2) when used to test hypotheses in two-way contingency tables. The clustering effect leads to unacceptably high Type I errors (Thomas and Rao (1987)) which do not satisfy the assumptions of multinomial sampling. Thomas and Rao(1987) found in a number of simulation studies that, under cluster sampling, the proportion of rejections of a correct null hypothesis (at a nominal level of significance of 5% using the uncorrected chi-square or likelihood ratio goodness-of-fit test statistic) becomes unacceptably large, even as large as 50%. Hence some adjustment is necessary so that misleading results are prevented. Various alternative test statistics as well as adjustments to the Pearson chi-square test statistic have been proposed, which take into account the design of the survey (Fellegi, 1980, Rao and Scott, 1981).

In order to allow for the simultaneous examination of all pairwise relations in a multidimensional contingency table, including the possibility of 3-factor and higher interactions, a log-linear model can be used. This model is most often used when no distinction is made between the factors (all are regarded as explanatory).

The approach used to analyze the multiway table is to scale the weighted totals, so that the sample size is maintained, and to use a correction factor (Fellegi, 1980, Rao and Scott, 1981) during the model selection phase. This correction factor adjusts the chi-square test statistic to compensate for the complexity of the survey design.

7.2 TEST STATISTICS FOR ANALYSING TWO-WAY AND MULTIWAY TABLES USING THE COVARIANCE MATRIX UNDER THE SAMPLE DESIGN

Koch, Freeman and Freeman(1975) noted that the weighted least squares method first advocated by Grizzle, Stamer, and Koch(1969)(GSK) method could be extended to the analysis of complex survey data. The GSK method describes how an underlying multinomial distribution is assumed for the frequencies in the contingency table when estimating the cell proportions and their covariance matrix. Large-sample Wald statistics, which have asymptotic chi-square distributions, are used to assess the significance of the models fitted to the cell proportions. Koch, Freeman and Freeman(1975) extended this theory by estimating the cell proportions and their covariance matrix under the sample design. The analysis then proceeds using the sample design estimates in place of the multinomial estimates.

Consider the linear model

$$P = X\beta + \epsilon$$

where ϵ is the error term, and the left hand side is the column vector of all the proportions computed for each cell in a contingency table. The proportions are formed by taking the numerators to be the weighted cell frequency estimates for each cell defined by the complete cross of all the variables specified, and the denominator to be the grand total over all these variables. The proportions are arranged in a long vector $\hat{P} = \{P_1, P_2, \ldots, P_{IJK}\}$ and the variance-covariance matrix of $\hat{P}$ is estimated on the basis of the specified survey design. A design matrix (X) based on the model is constructed to satisfy the relationship

$$E(\hat{P}) = X\beta$$

For the loglinear model each element of $\hat{P}$ is replaced by $\log(P_{ijk})$. The variance-covariance matrix of these quantities, V, is estimated from the survey data. The weighted least squares estimator of the coefficients $\hat{\beta}$, is given by

$$\boldsymbol{\beta} = [X'V^{-1}X]^{-1}X'V^{-1}\hat{P} \qquad (7.2.1)$$

and the variance covariance matrix of $\hat{\beta}$ by

$$\hat{V}(\beta) = [X'V^{-1}X]^{-1} \qquad (7.2.2)$$

The tests of hypothesis are based on $\hat{\beta}$ and $\hat{V}(\hat{\beta})$ using the theory below.

HYPOTHESIS TESTING - COMPUTATIONAL APPROACH

There are a variety of test statistics for testing the null hypothesis

H_o: HP = 0 versus H_a: HP $\neq$ 0

where P is the column vector of population proportions and H is the hypothesis matrix of full row rank c.

7.2.1 Wald Chi-Square Statistic

Weighted least squares methods based on the Wald chi-square statistic (Koch, Freeman and Freeman 1975) have been extensively used to analyze survey data and it is given by

$$Q = [H\hat{P}]'[H\hat{V}H']^{-1}[H\hat{P}]$$

Under H_o, Q is asymptotically distributed as a chi-square random variable with c degrees of freedom.

Simulation studies by Thomas and Rao(1984 and 1985) indicate that the Wald chi-square test statistic is too liberal when the rank of H is large relative to the degrees of freedom associated with

$\hat{V}$. That is, Q ejects the null hypothesis more often than the established nominal level when the null hypothesis is true.

7.2.2 Wald F Statistic

The Wald F statistic is obtained by dividing Q by c, the rank of H. That is

$$F_W = Q/c$$

Under H_o, F_W is asymptotically distributed as an F random variable with c and e degrees of freedom, where e is the degrees of freedom associated with $\hat{V}$. For multistage designs e is usually taken to be the number of PSU's minus the number of first stage strata. When c-e is not large, F_W may reject H_o more often than the nominal level (see Thomas and Rao (1984)).

7.2.3 Adjusted Wald F Statistic

The adjusted Wald F statistic was proposed by Folsom(1974). It compensates for the liberal nature of the Wald statistic by approximating the number of degrees of freedom associated with $\hat{V}$ and using the more conservative Hotelling T^2 distribution for Q. The F transformed Wald statistic is given by

$$F_{ADJWF} = \frac{c - e + 1}{ce} Q$$

Under the null hypothesis , F_{ADJWF} can be compared to an F distribution with e and c-e+1 degrees of freedom (Fellegi, 1980). This simple transformation improves the performance of the test (Korn and Graubard (1990)).

7.2.4 Satterthwaite Adjusted Chi-Square Statistic

For categorical data analysis Rao and Scott's (1981) Satterthwaite correction is based on the Pearson chi-square statistic, which assumes multinomial sampling, and adjusts the test statistic to reflect the impact of the clustered sample design. Rao and Scott showed that

$$\chi^2 \sim \sum_i \lambda_i w_i$$

where the λ's are the eigenvalues of Rao and Scott's generalized design effect matrix and the w's are independent chi-square random variables with one degree of freedom. The sum of the diagonal elements of the generalised design effect matrix is equal to the sum of the generalised design effects, the average of which is equal to the first-order Rao-Scott adjustment (see Section 7.6.1). The Rao-Scott $\bar{\lambda}$ corrected statistic adjusts the Pearson chi-square by dividing by $\bar{\lambda}$, the average of the eigenvalues. The Rao-Scott Satterthwaite correction uses Satterthwaites (Satterthwaite(1946)) approach to account for the variability in the eigenvalues.

The Satterthwaite adjusted chi-square statistic (Rao and Scott 1981) takes the form

$$\chi^2{}_s = \frac{Q^*}{(\bar{\lambda}(1 + a^2))} \quad where$$
$$Q^* = [H\hat{P}]'[HSH']^{-1}[H\hat{P}]$$

S is the estimated variance covariance matrix under multinomial sampling, and a is the coefficient of variation of the λ's. The degrees of freedom are also adjusted by dividing by $(1 + a^2)$. That is, $\chi^2{}_s$ has degrees of freedom c^* where
$c^* = c/(1 + a^2)$. In order to compute $\chi^2{}_s$, the generalized design effect matrix, D, is computed using the formula

$$D = [HSH']^{-1}[H\hat{V}H']$$

where $\hat{V}$ is the estimated variance covariance matrix. In the above formula for χ^{*}_{s}, $\bar{\lambda}$ represents the average of the eigenvalues of D. In their simulation studies, Thomas and Rao(1987) found that the Rao-Scott Satterthwaite corrected chi-square statistic had the best overall properties when both power and significance level were considered. This test maintained its power and stated significance level better than any of the alternatives across the combinations of the designs considered.

7.3 USING A WALD STATISTIC TO TEST FOR HOMOGENEITY AND NO INTERACTION IN A TWO-WAY TABLE

Under simple random sampling, testing for homogeneity and independence in a two-way table are treated in the same way. This is not the case under a complex survey design (Stoker et al. 1997). The test of homogeneity of two populations implies that the two (sub-)populations form geographic classes (see Section 6.6) where the sample elements are classified into mutually exclusive and exhaustive categories with fixed marginal totals. For example, from the NH&NS, place of residence is such an example. A PSU either belongs to a rural or urban sub-population. The objective of the test of homogeneity in this case is to test whether the vectors of population proportions are the same in the rural and urban sub-populations. The null-hypothesis in this case is

H_o: $p_{11} = p_{21}; p_{12} = p_{22}; \ldots; p_{1(c-1)} = p_{2(c-1)}$

or H_o: $\underline{p}_1 = \underline{p}_2$

where $\underline{p}_1 = (p_{11}\ p_{12}\ p_{13}\ \cdots\ p_{1(c-1)})$ and $\underline{p}_2 = (p_{21}\ p_{22}\ p_{23}\ \cdots\ p_{2(c-1)})$

and c is the number of categories.

and the Wald statistic is

$$Q = (\hat{\underline{p}}_1 - \hat{\underline{p}}_2)'(\hat{V}(\hat{\underline{p}}_1) + \hat{V}(\hat{\underline{p}}_2))^{-1}(\hat{\underline{p}}_1 - \hat{\underline{p}}_2) \qquad (7.3.1)$$

When the null hypothesis is true Q is distributed as a chi-squared statistic with c-1 degrees of freedom.

The above approach (using a Wald chi-square statistic) does not apply to a complex survey design where the interaction between two categorical variables is tested. Interaction can be defined as a differing response in one variable depending on the level of another variable. In this case the two categorical variables form either cross classes or mixed classes (Stoker et al. 1997) (see Section 6.6). The data are assumed to be drawn from a single population with no fixed marginal totals.

Using the Wald statistic, a test for no interaction in the case of a two-way table from a complex survey design, is shown below.

Let $\delta_i(r,c)$ reference the cell corresponding to the r-th row and c-th column in a two-way cross-classification with r=1,...,R and c=1,...,C, and let

$$\delta_i(r,c) = \begin{cases} 1 \text{ if the ith sample member is in the } (r,c) \text{ cell} \\ 0 \text{ otherwise} \end{cases}$$

Also let

$$\delta_i(r) = \begin{cases} 1 \textit{ if ith sample member is in the r-th row} \\ 0 \textit{ otherwise} \end{cases}$$

and

$$\delta_i(c) = \begin{cases} 1 \textit{ if ith sample member is in the c-th column} \\ 0 \textit{ otherwise} \end{cases}$$

Then the total estimate for the (r,c) cell is given by

$$\hat{N}_{rc} = \sum_{i \in S} \delta_i(r,c)\, w_i$$

where w_i is the weight of the i-th sample member

A test for independence of the row and column variables is based on the following statistic for an RxC table (R= number of rows, C=number of columns)

$$\hat{Y}_{rc} = \frac{\hat{N}_{rc} - \hat{N}_{r+}\hat{N}_{+c}}{\hat{N}_{++}} \qquad (7.3.2)$$

which is the observed minus expected values (SUDAAN(1995)). Let $\hat{Y}$ denote the (R-1)(C-1) vector of elements $\hat{Y}_{rc}$. The variance-covariance of $\hat{Y}$ can be estimated as

$$\hat{V}(\hat{Y}) = H\hat{V}(\hat{N})H'$$

where

$$H = \frac{\partial\hat{Y}}{\partial\hat{N}'}$$

The matrix H is of dimension (R-1)(C-1) by RC and consists of partial derivatives of the elements of $\hat{Y}$ with respect to the elements of $\hat{N}$. For r=1,...,R-1, c=1,...,C-1, k=1,...,R, and l=1,...,C, the (rc,kl) element is equal to

$$\frac{\partial\hat{Y}_{rc}}{\partial\hat{N}_{kl}}$$

and can be calculated according to the following formulas,

$$\frac{\partial \hat{Y}_{rc}}{\partial \hat{N}_{rc}} = 1 - \frac{(\hat{N}_{r+} + \hat{N}_{+c} - \frac{\hat{N}_{r+}\hat{N}_{+c}}{\hat{N}_{++}})}{\hat{N}_{++}}$$

$$\frac{\partial \hat{Y}_{rc}}{\partial \hat{N}_{kc}} = \frac{(-\hat{N}_{r+} + \frac{\hat{N}_{r+}\hat{N}_{+c}}{\hat{N}_{++}})}{\hat{N}_{++}} \qquad k \neq r \qquad (7.3.3)$$

$$\frac{\partial \hat{Y}_{rc}}{\partial \hat{N}_{rl}} = \frac{(-\hat{N}_{+c} + \frac{\hat{N}_{r+}\hat{N}_{+c}}{\hat{N}_{++}})}{\hat{N}_{++}} \qquad l \neq c$$

$$\frac{\partial \hat{Y}_{rc}}{\partial \hat{N}_{kl}} = \frac{\hat{N}_{r+}\hat{N}_{+c}}{\hat{N}^2{}_{++}} \qquad k \neq r,\ l \neq c$$

for testing for no interaction in a two-way table the Wald statistic is

$$Q = \hat{Y}'[H\hat{V}(\hat{N})H']^{-1}\hat{Y} \qquad (7.3.4)$$

When the null hypothesis is true, Q is distributed as a chi-squared statistic with (R-1)(C-1) degrees of freedom. The p-value for the chi-square statistic is based on Q/(R-1)(C-1) which is compared to the F distribution with (R-1)(C-1) and e degrees of freedom, where e is the number of PSU's minus the number of strata.

7.4 FITTING AND TESTING THE LOG-LINEAR MODEL WITHOUT COMPENSATING FOR A WEIGHTED COMPLEX SURVEY DESIGN

Below is a description of the theory behind the log-linear model, assuming that the sample values are independent observations from the same population and that the selection probabilities of the sample values are equal, i.e. the sample is self-weighting. Once this basic theory has been described, techniques to compensate for a complex survey design are described. A log-linear model is then fitted to the child nutrition data (see Chapter 5) from the NH&NS, and some interesting results are reported.

7.4.1 THE LOG-LINEAR MODEL FOR A THREE-WAY TABLE

In a two-way table the test for independence between row and column factors is equivalent to testing the fit of the model:

$$E(f_{ij}) = \alpha_i \beta_j \qquad (7.4.1)$$

where the expected cell frequency $E(f_{ij})$ is the product of two terms, one of which depends only on the row (α_i) that the frequency appears in, and the other on the column (β_j) that the frequency appears in. In a higher way table, for example a three way IxJxK contingency table, the expected cell frequency may in general be expressed as the product of several terms, each representing a main effect or interaction. Since the logarithms of a product of terms is the sum of the logarithms of the terms, the logarithm of the expected frequency can be expressed as a linear model. Consider a three-way IxJxK contingency table, where the three indices pertain to categorical factors A,B,C, respectively. Let f_{ijk} be the observed frequency in cell (i,j,k) of the table. The log-linear model may then be written as

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC} + \lambda_{ijk}^{ABC} \qquad (7.4.2)$$

where $F_{ijk} = E(f_{ijk})$ is the expected value of the observed cell frequency (i,j,k) for the three-way model, θ = Log N, where N is the overall number of observations in the data set, and the λ's satisfy the constraints

$$\sum_i \lambda_i^A = 0, \ldots,$$
$$\sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = 0, \ldots, \qquad (7.4.3)$$
$$\sum_i \lambda_{ijk}^{ABC} = \sum_j \lambda_{ijk}^{ABC} = \sum_k \lambda_{ijk}^{ABC} = 0$$

The λ's are called the effects, with the superscripts indicating the factors to which the effects refer. The log-linear model given in (7.4.2) is a saturated model since it contains all possible effects. Usually the log-linear model is analyzed as a hierarchial model which means that a higher order effect cannot be present unless all lower order effects whose indices are subsets of the higher order effect are also included in the model. For example if λ^{AB} is present in the model, it means that effects λ^A and λ^B are also all present.

Letting $x_{ijk} = \ln F_{ijk}$, the main effects and interactions in the above model of a three-way table can be estimated using the 'delta' method (Lee (1977)) by

$$\hat{\lambda}_i^A = \overline{x}_{i\ldots} - \overline{x}_{\ldots\ldots}$$
$$\hat{\lambda}_{ij}^{AB} = \overline{x}_{ij..} - \overline{x}_{i\ldots} - \overline{x}_{.j..} - \overline{x}_{\ldots\ldots} \qquad (7.4.4)$$
$$\hat{\lambda}_{ijk}^{ABC} = \overline{x}_{ijk.} - \overline{x}_{ij..} - \overline{x}_{i.k.} - \overline{x}_{.jk.} + \overline{x}_{i\ldots} + \overline{x}_{.j..} + \overline{x}_{..k.} - \overline{x}_{\ldots\ldots}$$

where a period (.) indicates the mean of the omitted subscript.

Although this looks very much like an Analysis of Variance (ANOVA) model there is one major difference. In the ANOVA model one is trying to partition the variability in the response variable according to the levels of the factors. Here one is trying to describe structural relationships among the variables in the table i.e. one wants to determine which combinations of levels of the factors result in a higher cell count than expected, which a lower count, and which result in a cell count approximately that expected under the hypothesis of independence of the factors. When the parameter estimates cannot be obtained as described above in closed form by a simple expression, estimating the expected frequencies for some sub-models of the log-linear model requires iterative methods (see for example Agresti 1990).

The goodness-of-fit of the model can be tested using the usual Pearson goodness-of-fit chi-square statistic

$$\chi^2 = \sum_{ijk} \frac{(f_{ijk} - \hat{F}_{ijk})^2}{\hat{F}_{ijk}} \qquad (7.4.5)$$

where $\hat{F}$=F with estimated parameters from (7.4.4)

or the likelihood-ratio statistic

$$G^2 = 2\sum_{ijk} f_{ijk}\ln(\frac{f_{ijk}}{F_{ijk}}) \qquad (7.4.6)$$

Both are asymptotically distributed as chi-square, with n-p degrees of freedom, where n is the number of cells, and p is the number of independent parameters estimated.

There are some indications that the G^2 statistic requires slightly larger sample sizes than the Pearson chi-square statistic to reach the same accuracy of p value (so that the Pearson chi-squqre statistic is more reliable for small sample sizes). However the likelihood ratio statistic G^2 has an advantage in that it is additive under partitioning for nested models. Two models M1 and M2 are said to be nested if all of the λ effects in M1 are a subset of the λ's contained in M2. The difference in G^2 between the two models is a test of the additional effects in M2 conditional on the effects in M1. This difference also has an asymptotic chi-square distribution with degrees of freedom equal to the number of parameters fitted to the two models. This property does not hold for the Pearson chi-square statistic. Therefore in the tests that follow G^2 is used where a comparison is made of nested models.

7.4.2 MODEL SELECTION

The more variables that are included in the model, the better the model fits. However, the more terms there are in the model, the more difficult is the interpretation. One therefore aims for the simplest model that is adequate. The following variable selection methods can be used:

The **forward selection method** starts with no variables in the model. Initially the individual terms are considered, and those that are significant are included, one by one. Terms involving pairwise interactions in the individual terms that have been included, are then considered for inclusion. Then one considers three-way interactions in terms for which both the individual and the pairwise interactions are in the model. This process continues until no additional terms can be added to the model.

The **backward selection method** starts with the saturated model, that is, with the full model including the interaction terms between all factors, plus all interactions between combinations of these variables. The model without the highest order interaction, but with all other interactions present, is then considered to see if the decrease in fit is significant or not. If not, the term with the highest order interaction is eliminated, and the subset model is examined to see if any other interaction can be omitted without giving rise to a significant lack of fit.

In the case of a four-way table, $x_{ijkl} = \ln F_{ijkl}$ and the log-linear model may then be written as

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC} + \lambda_{il}^{AD} + \lambda_{jl}^{BD} + \lambda_{kl}^{CD} + \lambda_{ijk}^{ABC} + \lambda_{ijl}^{ABD} + \lambda_{jkl}^{BCD} + \lambda_{ijkl}^{ABCD} \qquad (7.4.7)$$

7.5 ADJUSTMENTS FOR WEIGHTED DATA

It has been remarked that many researchers use one of two strategies to analyze weighted data: (1) analyze the unweighted frequencies by simply ignoring the weighting feature altogether, or (2) analyze weighted frequencies as if they were obtained from a data set without any weighting (Clogg and Eliason (1987)). Using either of these strategies, could lead one to incorrect inferences, parameter estimates can be biased, fit statistics can be misleading, and standard errors can be incorrect. The approach suggested by Clogg and Eliason (1987), which is the approach that is applied in this chapter, is to rescale the weights so that

$$\sum_h w_h = n \qquad (7.5.1)$$

where w_h is the weight of the h-th observation. This ensures that the sample size is preserved and the weighted frequencies are analyzed as if they are obtained from a sample without weighting features (see Clogg and Eliason(1987)).

Below is an example of a two-way table for the variables gender and stunting from the NH&NS. The scaled totals are calculated using the sample totals and the weighted proportions.

Table 7.3 Sample frequencies and scaled frequencies of stunted children by sex in the Lesotho National Household Health and Nutrition Survey

sample totals weighted proportion scaled totals design effect	female	male	Total
stunted	959 0.115 985 1.924	1041 0.124 1062 2.314	2000 0.239 2047
not stunted	3355 0.393 3367 1.815	3213 0.367 3144 1.753	6568 0.760 6511
Total	4314 0.503 4352	4254 0.497 4206	8568 1.000 8558

7.6 CORRECTION FACTORS TO ADJUST THE PEARSON CHI-SQUARE TEST STATISTIC FOR A TWO-WAY TABLE AND THE LIKELIHOOD RATIO GOODNESS-OF-FIT TEST STATISTIC FOR A LOG-LINEAR MODEL

The Pearson chi-square statistic for a two-way table is given by

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed number in cell (i,j), $E_{ij}=r_ic_j/n$ is the expected number in cell (i,j) and n is the number of cells in the two-way table.

To test for no interaction between two categorical variables the null hypothesis is

$H_o : p_{ij} = p_{i+} \cdot p_{+j}$ for al i,j

where p_{i+} is the marginal proportion in row i and p_{+j} is the marginal proportion in column j.

To test for homogeneity between two populations the null hypothesis is

$H_o: p_{11} = p_{21}; p_{12} = p_{22}; \ldots; p_{1(c-1)} = p_{2(c-1)}$

where c is the number of categories.

Since Pearsons chi-square statistic tends to become too large in cluster sampling, various modifications to this goodness-of-fit test statistic have been suggested to obtain more reliable results in cluster sampling. Fellegi and Rao and Scott both modify the chi-square test statistic by dividing it by a correction factor which is a function of the design effect values of the cells of the contingency table.

Rao and Scott (1981) stress the importance of assessing the impact of the survey design on the Pearson statistic by including simple correction factors instead of using the full covariance matrix (using the Wald Statistic in Section 7.3) for the following two reasons:
" a) In the preliminary analysis of survey data, the Pearson criterion function is often used, because of its computational simplicity, to screen out a large number of two-way (or higher dimensional) tables at minimal cost to identify tables of interest; b) in secondary analysis from published reports containing multiway contingency tables, the researcher may not have access to the full covariance matrix."

7.6.1 RAO AND SCOTT'S $\hat{\lambda}$ CORRECTION

For the test of independence between two categorical variables this method makes the following corrections to the Pearson chi-square statistic (Rao and Scott 1981) :

$$\chi^2_c = \chi^2/\hat{\lambda}$$

which is compared to the $\chi^2_{(r-1)(c-1)}$ distribution, where

$$\hat{\lambda} = (rc - 1)^{-1}[\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{d}_{ij}(1 - \hat{p}_{i+}\hat{p}_{+j})] \qquad (7.6.1)$$

and $\hat{d}_{ij}$ is the design effect of the ij-th cell.

For the test of homogeneity between two segregated populations the Pearson chi-square statistic is also divided by the correction factor so that:

$$\chi^2_c = \chi^2/\hat{\lambda}$$

where

$$\hat{\lambda} = \frac{1}{c-1}\sum_{i=1}^{2}(1 - \frac{\hat{n}_i}{\hat{n}_1 + \hat{n}_2})\sum_{j=1}^{c}\frac{\hat{p}_{ij}}{\hat{p}_{+j}}(1 - \hat{p}_{ij})\,\hat{d}_{ij}$$

(7.6.2)

and $\hat{n}_1$ and $\hat{n}_2$ are the totals in the two segregated populations. In this case χ^2_c is compared to the $\chi^2_{(c-1)}$ distribution (Stoker et al. 1997).

7.6.2 FELLEGI'S CORRECTION FACTOR

In both the test of no interaction and homogeneity, Fellegi's approach corrects χ^2 by dividing it by the average design effect of the two-way table where

$$\bar{d} = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{d}_{ij} \qquad (7.6.3)$$

i.e.

$\chi^2_F = \chi^2/\bar{d}$

which is compared to the $\chi^2_{(r-1)(c-1)}$ distribution in the test of no interaction and to the $\chi^2_{(c-1)}$ distribution in the test of homogeneity. For a discussion of the general case of more than two geographic or segregated classes see Särndal et al. (1992) and Nathan (1988).

In the case of multiway tables where a log-linear model is fitted to the data to simultaneously assess associations between variables, the likelihood ratio statistic is also corrected for a complex survey design using Fellegi's correction factor. For example in the case of a three-way table (see Section 7.4.1)

$$\bar{d} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{d}_{ijk} \qquad (7.6.4)$$

and using Fellegi's correction factor

$G^2_F = G^2/\bar{d}$

which (from (7.4.5)) is compared to the $\chi^2_{(n-p)}$ distribution where n is the number of cells in the multiway table and p is the number of independent parameters being estimated. In the case of three-or more-way tables, computing Rao and Scott's $\hat{\lambda}$ correction factor becomes complicated, as it becomes dependent on the hypothesis being tested in the log-linear model.

7.6.3 EXAMPLES OF TWO-WAY TABLE ANALYSES FROM THE NH&NS USING DIFFERENT CORRECTION FACTORS

Stunting and gender in Table 7.3 can be considered to be more or less uniformly distributed across the population and hence across all PSU's. In this case a test of independence (no interaction) between stunting and gender is applicable.

In order to test for independence between stunting and gender the ordinary chi-square test statistic of the rXc contingency table (r=number of rows, c=number of columns) is calculated after adjusting the sample totals as described in Section 7.5. The cell and average design effect is calculated using the program STANDERR. Therafter Pearsons chi-square statistic is adjusted by Fellegi or Rao and Scott's correction factors.

Fellegi's correction factor (see equation 7.6.3) $\bar{d}$ is firstly included in the test statistic for Table 7.3, where

$$\bar{d} = 1/4[1.924 + 2.314 + 1.815 +1.713]$$

$$= 1.942$$

and χ^2 is equal to 7.90396 which is significant (compared to the χ^2_1 distribution) at the 0.005 level. The adjusted chi-square statistic is therefore

$$\chi^2_c = 7.904/1.942$$

$$= 4.07$$

which is significant at the 0.05 level of significance but not at the 0.01 or 0.005 levels.

Rao and Scott's correction factor (see equation 7.6.1) to test for no interaction in Table 7.3 is

$$\hat{\lambda} = (1/3)(\{1 - [(0.234)(0.503)]\}(1.924) + \{1 - [(0.234)(0.495)]\}(2.314) + \{1 - [(0.503)(0.766)]\}(1.815) + \{1 - [(0.495)(0.766)]\}(1.753))$$
$$= 1.9839$$

Therefore $\chi^2_c = \chi^2/1.9839 = 3.984$ which gives a similar result to that using Fellegi's correction factor.

In Section 6.6 it was explained that a variable such as place of residence (rural/urban) forms a segregated class in the population. Therefore a test of homogeneity is appropriate to test whether the proportion of stunted children in Table 7.4 are the same in the rural and urban areas of the population. From (equation (7.6.2)) the Pearson chi-square statistic adjusted by Fellegi's correction factor is

$$\chi^2_c = 21.5568/3.205$$
$$= 6.726 \quad (p=0.0095)$$

which in this case is compared to a $\chi^2_{(1)}$ distribution.

From equation (7.6.2), Rao and Scott's correction factor is

$$\hat{\lambda} = (1 - 7214/8568)[\{(0.2466/0.2396)(1 - 0.2466)(3.88)\} + \{(0.7534/0.7604)(1 - 0.7534)(3.92)\}] + (1 - 1354/8568)[\{(0.1720/0.2396)(1 - 0.1720)(1.92)\} + \{(0.8280/0.7604)(1 - 0.8280)(3.10)\}]$$
$$= 2.0763$$

and $\chi^2_c = 21.5568/2.0763$

$= 10.382$ $(p=0.00127)$

In the above example, Fellegi and Rao and Scott's adjustment produce the same conclusions regarding the two-way table in Table 7.4. (see Table 7.7 for a more detailed presentation and discussion of these results). The slightly larger p-value for Fellegi can be attributed to it's conservative nature (Thomas and Rao 1981).

Table 7.4 Sample frequencies and scaled frequencies of stunted children by place of residence in the Lesotho National Household Health and Nutrition Survey

sample totals weighted proportion scaled totals design effect	stunted	not stunted	Total
rural	1757 0.2466 1779 3.88	5457 0.7534 5435 3.92	7214 1.000
urban	243 0.1720 233 1.92	1111 0.8280 1121 3.10	1354 1.000
Total	2000 0.2396 2012	6568 0.7604 6556	8568

It is clear from the above two examples that when the complexity of the survey design is not accounted for, misleading results can be produced, since for example a stronger association between stunting and gender is indicated before Fellegi's and Rao and Scott's correction factors are used to adjust the chi-square statistic in Table 7.3. It is also important to use the appropriate test statistic and correction factor when testing for homogeneity or independence in a two-way table.

Rao and Thomas (1989) remark that empirical results from various large-scale United Kingdom surveys indicated that Fellegi's χ^2_F produce very conservative tests in general, resulting in a significant loss of power relative to Rao and Scott's χ^2_c. This may not be serious if the sample sizes are large (Rao and Scott (1981)) as in the case of the NH&NS. Since SUDAAN does not calculate the Rao and Scott correction factor, Fellegi's correction factor is preferred to Rao and Scott's correction factor in analysing multiway tables and is used in subsequent analyses.

When there is a significant association between two categorical variables, the form of the association can be investigated by means of the standardised residuals. Stoker et al. (1997) show, using the software package PC CARP, that in general, standardised residuals under simple random sampling are larger than standardised residuals for a complex design. Since I did not have access to PC CARP (which can calculate the design effect of a residual of any cell in a contingency table), I could not include this form of analysis. The residuals have implications for log-linear models too, which points to the importance of taking the design of the survey into consideration when analysing data from contingency tables. In all further analyses (see Section 7.7) where log-linear models are fitted, the complexity of the design is only accounted for in the model selection phase (see Section 7.4.2) where, as in the example above, the likelihood ratio statistic is divided by the average design effect from the contingency table.

It is clear from the above that there are certain limitations with SUDAAN in analysing two-way tables from a complex survey design:

- There is no test of homogeneity using a Wald statistic.
- SUDAAN does not calculate Fellegi's correction factor (The SAS program STANDERR was used).
- SUDAAN does not calculate either of Rao and Scott's adjustments i.e. when the covariance matrix under the sample design is used (Section 7.3.2) or when the design effect in each cell of the

contingency table is used (Section 7.6.1).

- SUDAAN does not calculate standardised residuals in analysing contingency tables. The two-way and three-way table analyses tha follow in this chapter and in Chapter 9 make use of Fellegi's correction factor technique (using STANDERR) and Wald's chi-square statistic (using SUDAAN) to test for independence and homogeneity.

7.7 SOME INTERESTING RESULTS FROM THE NH&NS

The techniques described in the previous Sections are applied to the child nutrition data to produce some interesting results. Tests of association for two-way tables are applied using Fellegi's correction factor and Wald's statistic. The indicators of malnutrition: height-for-age, weight-for-age, and weight-for-height are tested for association with each of the predictor variables age, sex, place of residence (rural/urban), survey round, and district. Also of interest to the Bureau of Statistics in Lesotho are the breakdown of certain two-way tables by age group. Graphical representations are provided in those cases. In Chapter 8 malnutrition is predicted by all the abovementioned variables in a single logistic regression model. Therefore in the following examples a log-linear model is fitted to a multiway table only in the case where all the variables are considered to be explanatory. As an example of such a case, the simultaneous associations of the variables breastfeeding, malnutrition, and education level of the mother are investigated.

The number, percentage, and 95% confidence intervals (see Chapter 6) in the following tables are based on the totals weighted to the population, taking into consideration the complexity of the survey design. Two chi-square statistics are presented, the first with Fellegi's correction factor (χ^2(F)) (Section 7.6), and the second the chi-square based on the Wald statistic (χ^2(W)) (Section 7.3). Since the Wald statistic uses the full covariance matrix of the proportions and it produces a stable result, and since Fellegi's correction factor is known to produce conservative results (Rao and Thomas(1989)), (χ^2(W)) will be preferred to (χ^2(F)) in drawing conclusions regarding the test of independence in the following tables.

Table 7.5 Number and percentage of children under 5 years old below -2 Z-scores by age group

	Height/age		Weight/age		Weight/height	
Age group (months)	Number	% (95% CI)	Number	% (95% CI)	Number	% (95% CI)
0-11	5348	12.4 (10.2-14.7)	2382	5.4 (4.0-6.9)	1572	3.8 (2.5-5.0)
12-23	13143	28.4 (25.6-31.2)	6417	13.7 (11.9-15.5)	2079	4.5 (3.1-5.8)
24-35	11111	23.9 (21.1-26.7)	6852	14.7 (12.7-16.7)	1604	3.4 (2.3-4.6)
36-47	12406	26.2 (23.6-28.9)	7348	15.7 (13.8-17.7)	2352	4.9 (3.5-6.4)
49-59	12311	29.7 (26.8-32.7)	5912	14.5 (12.3-16.8,	1274	3.1 (2.0-4.1;
Weighted population totals	224614	overall % = 24.2% (22.4-25.9)	224636	overall % = 12.7% (11.7-14.1)	224601	overall % = 3.9% (3.1-4.8)
	χ_4^2(F)=101.597 χ_4^2(W)=154.961	p=0.000 p=0.000	χ_4^2(F)=78.552 χ_4^2(W)=161.88	p=0.000 p=0.000	χ_4^2(F)=6.3835 χ_4^2(W)=10.212	p=0.1728 p=0.0453

In Table 7.5 a significant difference between the percentage of children falling below 2 Z-scores below the mean in the different age groups for the weight-for-age and height-for-age indices is observed, but not for the weight-for-height index. The effect of poor nutrition on stunting is clearly evident in the 0-11 and 12-23 month age groups. The 12-23 month age group is 16% higher than the 0-11 month age group and after that the percentage plateaus for the remaining age groups. It would seem that the significant association between age and height-for-age is mainly due to the low percentage in the 0-11 month age group. The low proportion (5.4%) of underweight children (low weight-for-age) in the 0-11 month age group also seems to be the reason for the overall significant differences between the age groups for this index. These low rates under 1 year are most likely due to the high prevalence and long duration of breastfeeding which would tend to postpone the onset of stunting and low weight in children. With respect to low weight-for-height, it should be borne in mind that the overall proportion below two Z-scores only exceeds the normal population by 1.4 percent.

Table 7.6 Number and percentage of children below -2 Z-scores, by gender

Gender		Height/age		Weight/age		Weight/height	
		N*	% (95% CI)	N*	% (95% CI)	N*	% (95% CI)
All children under 5 years	Female	25152	22.8 (20.9-24.7)	13579	12.3 (11.0-13.5)	4076	3.7 (2.7-4.7)
	Male	29167	25.5 (23.5-27.5)	15360	13.5 (12.0-14.9)	4806	4.2 (3.2-5.2)
		χ_1^2(F)=4.125 p=0.042 χ_1^2(W)=9.67 p=0.003		χ_1^2(F)=1.904 p=0.167 χ_1^2(W)=6.14 p=0.015		χ_1^2(F)=0.645 p=0.422 χ_1^2(W)=1.94 p=0.168	
Under 24 months only	Female	7856	18.0 (15.7-20.2)	3278	7.4 (6.1-8.6)	1527	3.5 (2.1-4.9)
	Male	10635	23.3 (20.8-25.7)	5521	11.9 (10.0-13.8)	2124	4.7 (3.4-6.0)
		χ_1^2(F)=7.180 p=0.0074 χ_1^2(W)=14.29 p=0.0003		χ_1^2(F)=15.804 p=0.000 χ_1^2(W)=36.081 p=0.000		χ_1^2(F)=1.541 p=0.214 χ_1^2(W)=4.481 p=0.037	

* N is the weighted number in the population

When considering all children under 5 years in Table 7.6, there is significantly more stunting in males than in females but no strongly significant differences appear in the weight-for-age and weight-for-height indices. Figures 7.1, 7.2, and 7.3, show the three indices taking age and sex into consideration (In Figures 7.1 and 7.3, where height is measured, the discontinuity at 24 months mentioned earlier in Section 5.2.3 of Chapter 5 can be observed.). It is interesting to note that most of the gender difference in stunting is occurring below 24 months, and that this pattern is reproduced in the weight-for-age and weight-for-height plots. When analysed separately in two age intervals (under 24 months and over 24 months) none of the indices show a significant difference above 24 months but as shown in Table 7.6, below 24 months, malnutrition measured by height-for-age and weight-for-age is significantly greater in boys than in girls. In summary, stunting reaches a plateau around 26% in both sexes (see Figure 7.1), but more rapidly in boys. A similar trend is observed for weight-for-age in Figure 7.2. Of interest in Figure 7.2 is that males are significantly more underweight than females at the 0-11 months (p=0.0067) and 12-23 months (p=0.0001) age intervals. Thereafter no significant differences are evident for the age groups above 24 months. No significant differences exist between males and females for the acutely malnourished in any of the age groups (see Figure 7.3).

Figure 7.1 Percentages of children below 2 Z-scores with respect to height-for-age, plotted against age and sex

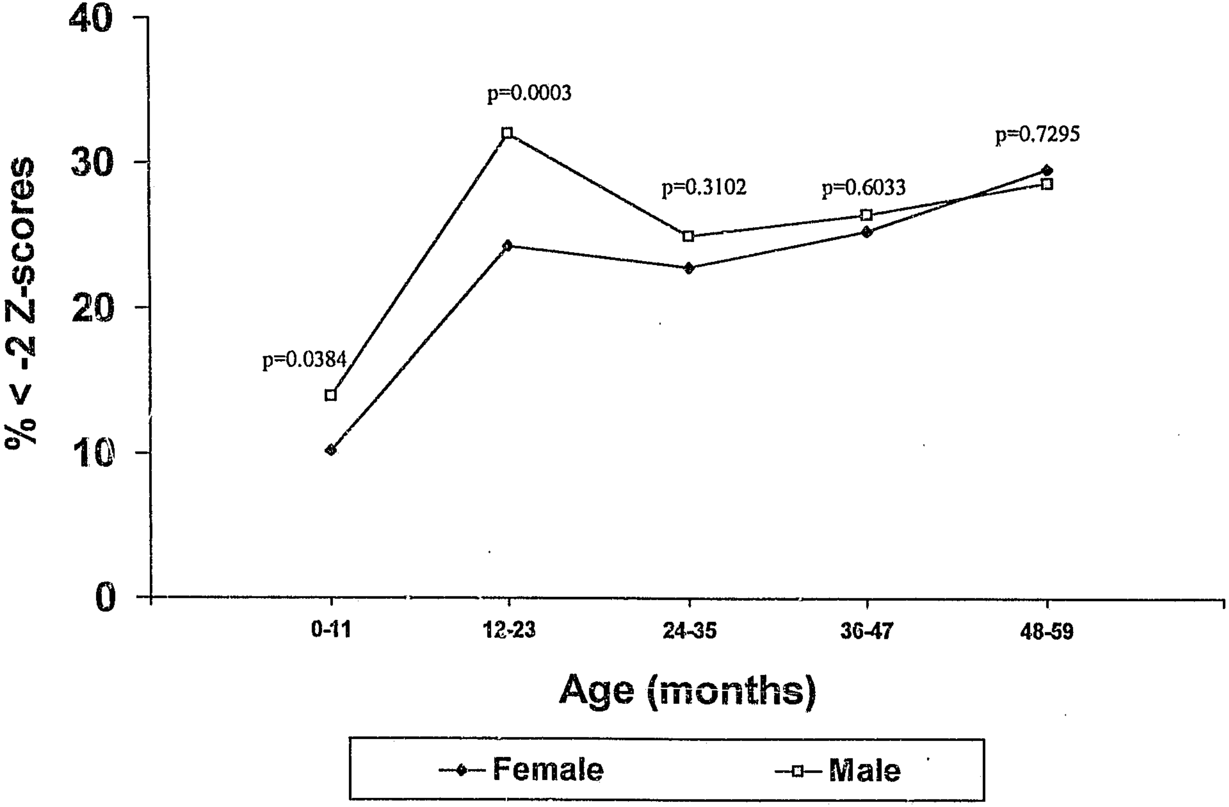


Figure 7.2 Percentages of children below 2 Z-scores with respect to weight-for-age, plotted against age and sex

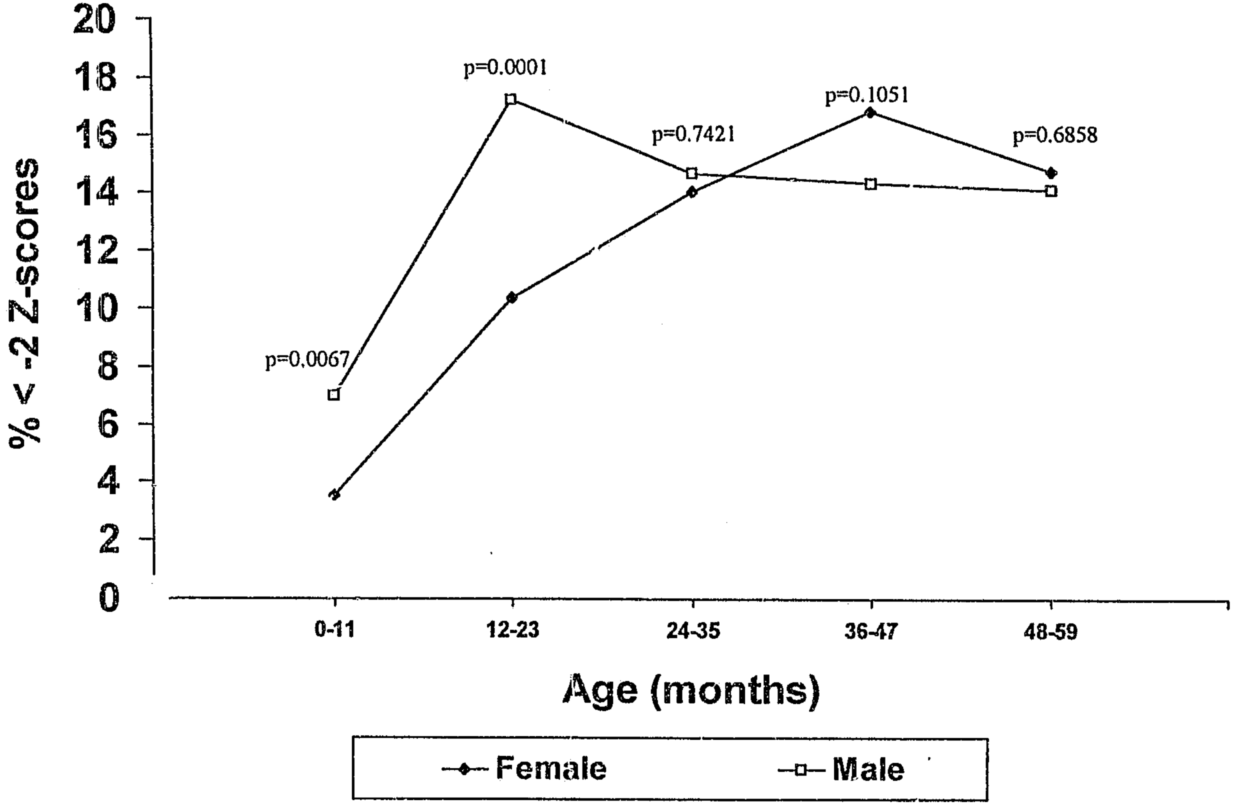


Figure 7.3 Percentages of children below 2 Z-scores with respect to weight-for-height, plotted against age and sex

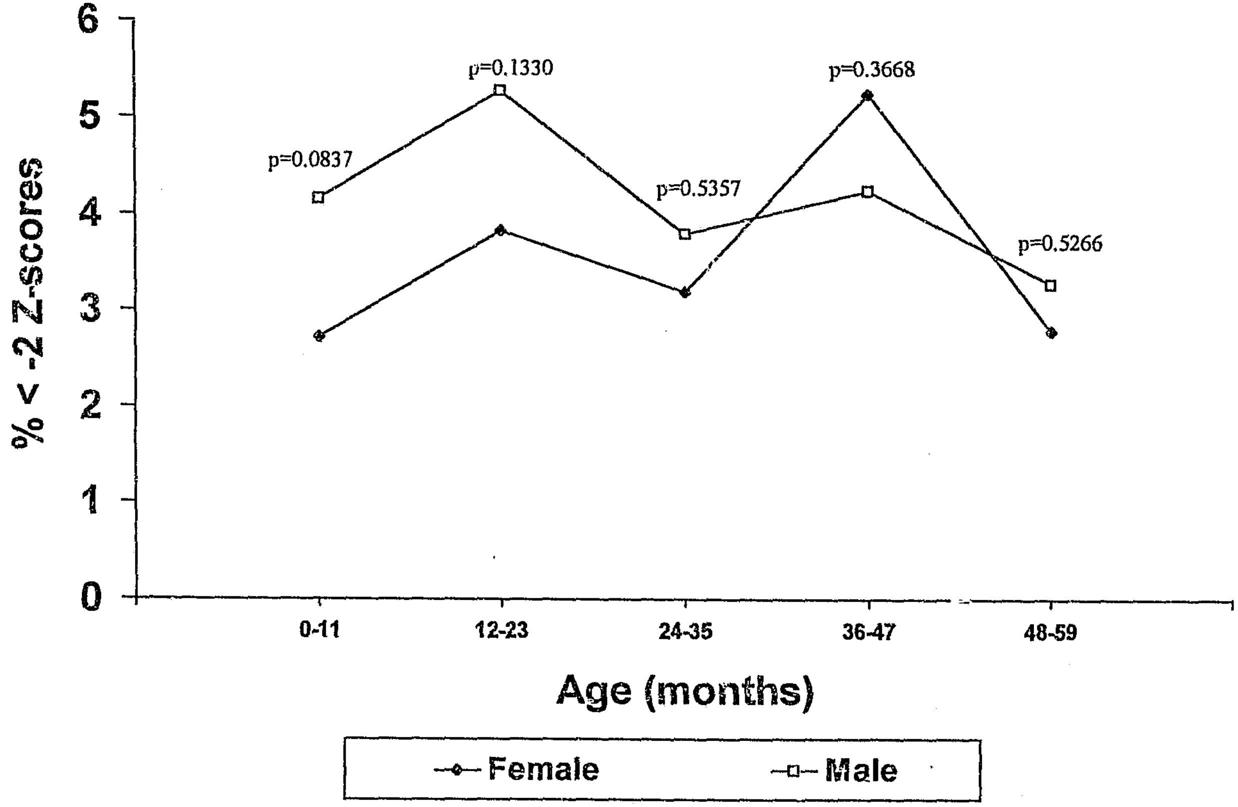


Table 7.7 Number and percentage of children under 5 years old, below -2 Z-scores, by round

	Height/age		Weight/age		Weight/height	
Round	Number	% (95% CI)	Number	% (95% CI)	Number	% (95% CI)
June 1988	54203	24.4 (21.4-27.3)	32041	14.4 (12.6-16.2)	13551	6.1 (4.3-7.8)
Nov 1988	56676	25.2 (22.2-28.2)	30362	13.5 (11.8-15.3)	9671	4.3 (2.9-5.7)
Feb 1989	51038	22.5 (20.0-24.9)	24952	11.0 (9.5 -12.6)	5217	2.3 (1.6-3.0)
	χ_2^2(F)=3.3272	p=0.1895	χ_2^2(F)=14.5025	p=0.0007	χ_2^2(F)=30.251	p=0.0000
	χ_2^2(W)=5.4333	p=0.0662	χ_2^2(W)=16.3310	p=0.0003	χ_2^2(W)=33.431	p=0.0000

The analysis displayed in Table 7.7 is performed on the separate rounds, where each round is treated as a separate survey with its own set of weights. Although there are no significant differences between the rounds in the case of stunting i.e. height-for-age, for the indices that reflect short term changes in nutritional status (weight-for-age and weight-for-height), a significant decrease is observed with the highest proportion of malnourished children found in the cold months of June, whilst the lowest proportion is observed in the summer month of February. The trends are also well illustrated in Figure 7.4.

This result is unexpected since during the winter months of June where harvesting takes place and food is more plentiful, the proportion of the acutely malnourished should be less than in the rainy summer months of February (Huss-Ashmore (1989)). Other factors such as source of income from migrant workers, breastfeeding patterns (see Chapter 8), or drought periods prior to the survey, need to be investigated to explain these results.

FIGURE 7.4 Percentage of children below −2 Z−scores with respect to height−for−age, weight−for−age and weight−for−height, plotted against round

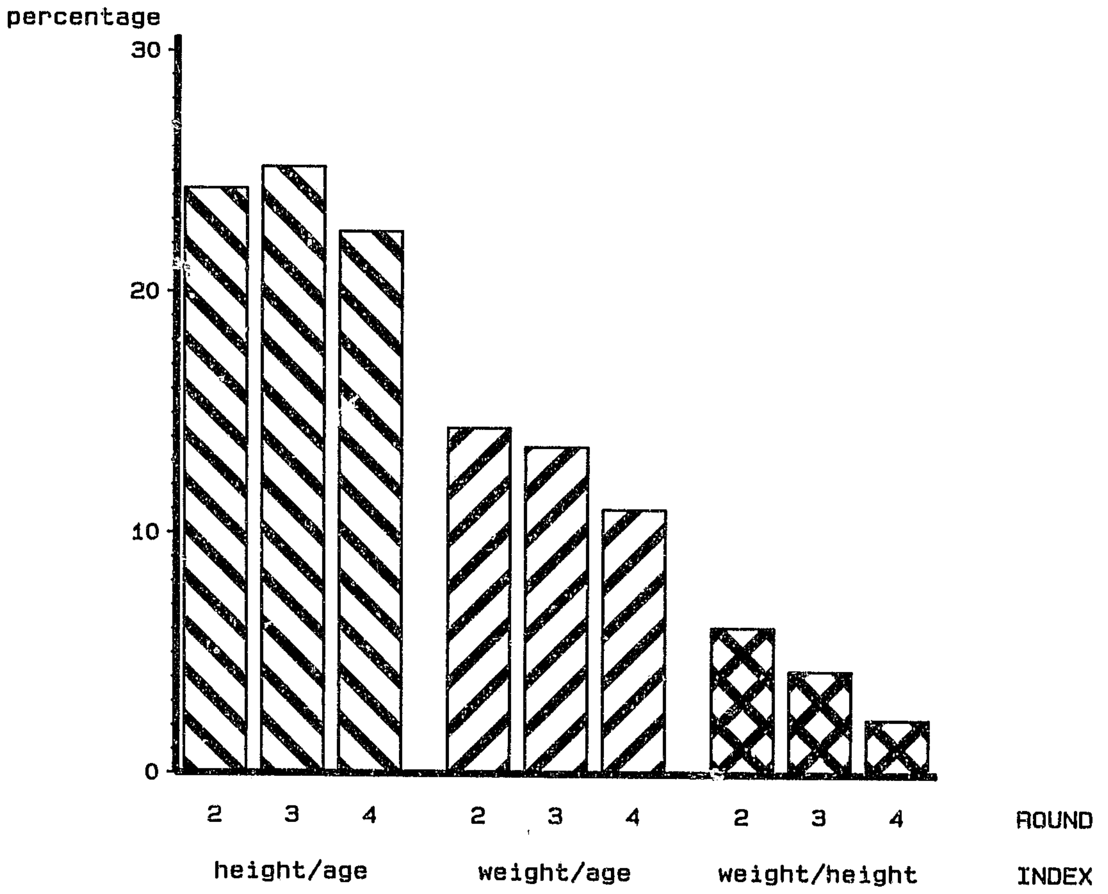


Table 7.8 Number and percentage of children under 5 years old below -2 Z-scores by place of residence

	Height/age		Weight/age		Weight/height	
	Number	% (95% CI)	Number	% (95% CI)	Number	% (95% CI)
Rural	50685	24.7 (22.8-26.7)	27097	13.3 (12.0-14.6)	8023	3.9 (3.0-4.9)
Urban	3634	17.2 (14.4-20.0)	1843	8.8 (6.3-11.4)	859	4.1 (2.3-5.9)
	$\chi_1^2(F)=6.726$	$p=0.0013$	$\chi_1^2(F)=5.865$	$p=0.0154$	$\chi_1^2(F)=0.015$	$p=0.9025$

The rural areas have a significantly higher proportion of stunted and underweight children than the urban areas. The plots of these data by the five age groups can be seen in Figures 7.5, 7.6, and 7.7. In Figures 7.5 and 7.6 (which refers to height-for-age and weight-for-age) the proportions malnourished are shown to be generally higher in the rural than urban areas in all the age groups. For height-for-age (Figure 7.5) the differences are significant at the 0-11 month (p=0.0005), 36-47 month (p=0.0022) and 48-59 (p=0.0016) month age intervals. For weight-for-age (Figure 7.6) the differences are significant in the same age groups: 0-11 months (p=0.045), 36-47 months (p=0.0030) and 48-59 (p=0.041). Figure 7.7 indicates there are significantly higher acutely malnourished children in the rural areas than in the urban areas in only the 24-35 month age interval. No significant differences appear in any of the other age groups.

The above analysis includes only a test of homogeneity (using Fellegi's correction factor) since rural and urban are two mutually exclusive and exhaustive populations and therefore (from Section 7.6) it would not be appropriate to test for independence in this case.

Figure 7.5 Percentages of children below 2 Z-scores with respect to height-for-age, plotted against age and place of residence

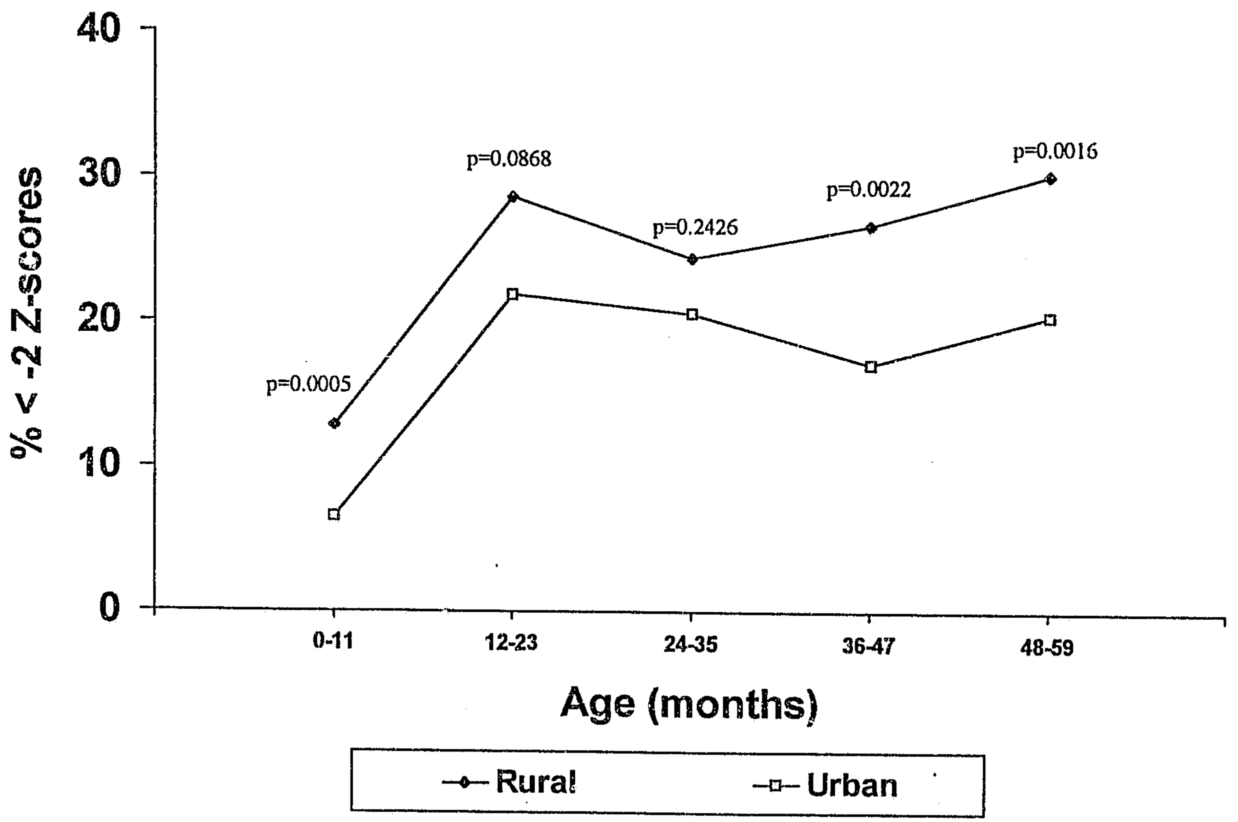
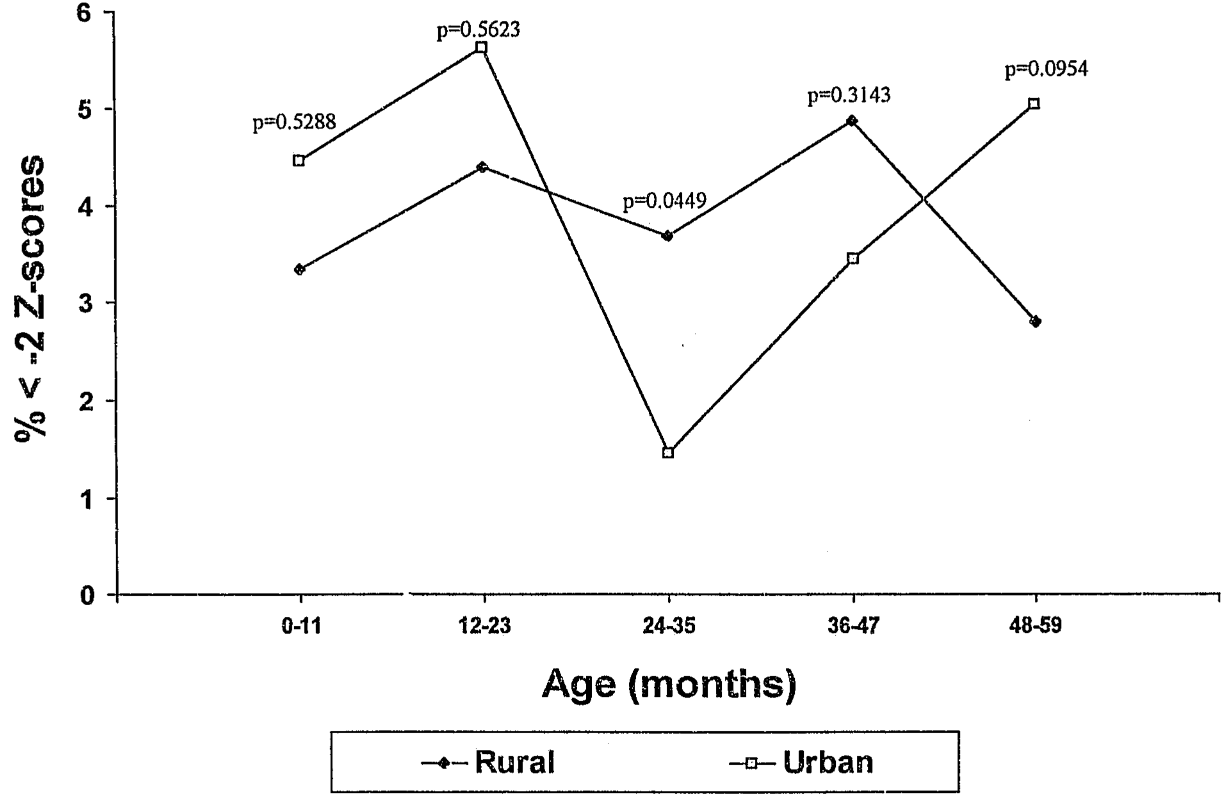


Figure 7.7 Percentages of children below 2 Z-scores with respect to weight-for-height, plotted against age and place of residence



The 10 districts of Lesotho in Table 7.9 indicate substantial variation in their levels of childhood malnutrition, with the more rural districts such as Thaba-Tseka having double the level of underweight and 50% more stunting than the more urbanised districts such as Butha-Buthe and Maseru (see Figure 7.8.).

Table 7.9 Number and percentage of children under 5 years old, below -2 Z-scores, by district

	Height/age		Weight/age		Weight/height	
District	N*	% (95% CI)	N*	% (95% CI)	N*	% (95% CI)
Butha-Buthe 1	2903	20.3 (13.5-26.9)	1266	8.8 (8.3-9.3)	510	3.5 (0.7-6.4)
Leribe 2	9386	25.5 (20.0-31.0)	4116	11.2(8.8-13.5)	949	2.6 (1.6-3.5)
Berea 3	4404	20.7 (17.0-24.4)	2381	11.2(9.8-12.6)	714	3.3 (1.8-4.8)
Maseru 4	10028	22.3 (17.9-26.7)	6336	14.1(10.1-18.1)	2611	5.8 (2.6-9.0)
Mafeteng 5	6497	22.0 (18.3-25.7)	3755	12.7(9.9-15.5)	833	2.8 (1.2-4.4)
Mohale Hoek 6	6401	26.0 (21.8-30.3)	3392	13.8(10.9-16.7)	927	3.8 (1.2-6.4)
Quthing 7	4318	25.0 (16.5-33.5)	2267	13.1(7.9-18.3)	607	3.5 (1.1-6.0)
Quacha Nek 8	2442	25.4 (23.8-26.9)	1053	10.8(8.8-12.9)	334	3.5 (0.0-7.1)
Mokhotlong 9	3290	27.6 (19.2-36.1)	1995	16.8(10.0-23.6)	650	5.5(0.0-11.2)
Thaba-Tseka 10	4650	32.5 (29.7-35.3)	2377	16.6(12.2-21.1)	746	5.2 (3.3-7.1)
	χ_1^2(F)=10.047	p=0.3467	χ_1^2(F)=8.749	p=0.4608	χ_1^2(F)=7.032	p=0.634
	χ_1^2(W)=34.510	p=0.0005	χ_1^2(W)=28.50	p=0.0026	χ_1^2(W)=11.93	p=0.237

It is clearly shown from all the above analyses that Fellegi's χ^2(F) is consistently smaller than χ^2(W), thus supporting the empirical results of Rao and Thomas(1989) where Fellegi's correction factor was shown to produce conservative results.

Of interest to the Lesotho BOS are the associations between breastfeeding, malnutrition, and age (below and above 24 months). In order to determine the associations simultaneously, a log-linear model is applied to the data. The SAS program STANDERR that has been written for this thesis (from 7.6.1) using Fellegi's correction factor is used for the analysis. The results generated by SUDAAN for a log-linear analysis were not quite clear to interpret from the printouts or the manuals. Therefore it has been decided rather to apply the correction factor technique using STANDERR to illustrate the log-linear model below. The data from the adjusted sample (see Section 7.5) given in Table 7.10 is used for the analysis.

Table 7.10 Frequency Table used to fit a Log-Linear Model to determine associations between breastfeeding, age, and malnutrition

		Age < 24 months		Age >= 24 months	
		malnourished	not malnourished	malnourished	not malnourished
Still breastfed	Sample size	287	2779	91	297
	Adj. sample	285	2771	95	293
	Weighted	5705	55411	1906	5854
	Design Effect	1.58	1.55	1.14	1.67
Not breastfeeding anymore	Sample size	68	516	659	4245
	Adj. sample	67	507	674	4249
	Weighted	1337	10133	13482	84968
	Design Effect	1.22	1.77	1.80	1.76

The saturated model for the above table is the same as that given in equation (7.4.2), where A refers to the age group of the child, B indicates whether the child is still breastfed or not, and C indicates whether the child is malnourished or not.

The backward model selection technique described in Section 7.4.2, is used to determine the appropriate log-linear model for the above example.

Table 7.11 Probability that the saturated model less the three-way interaction term is adequate

Effect	DF	G^2	P-value
base model	0	0.0000	1.000
ABC	1	23.66	0.000

The base model in Table 7.11 is the saturated model which fits the data perfectly, and therefore there are no degrees of freedom for testing this model. The second line is the model without the three-way interaction term ABC. The incremental effect of ABC is examined by comparing the difference between the G^2_{adj} (where G^2_{adj} = G^2/(average deff) = 23.66/1.56 = 15.17) for the saturated model and for the model without the ABC term. Thus the value for G^2_{adj} for the incremental effect is 15.17 - 0.0000 = 15.17, which is highly significant (p-value = 0.000). Therefore the three-way interaction does make a significant contribution to the model and cannot be eliminated from the model. As a result the saturated model is fitted to this data and therefore the standardised residuals are all equal to zero. The coefficients for the fitted model are given in Table 7.12.

Table 7.12 **Estimated model coefficients resulting from the selected model**

Factor	Coefficient	Std. Err.	Z-Value	95% CI
Constant	6.13930	0.02386	257.338	(6.092;6.186)
$\hat{\lambda}_1^A$	-0.13613	0.02386	-5.70622	(-0.183;-0.089)
$\hat{\lambda}_2^A$	0.13613	0.02386	5.70622	(0.089;0.183)
$\hat{\lambda}_1^B$	-0.18592	0.02386	-7.79319	(-0.233;-0.139)
$\hat{\lambda}_1^B$	0.18592	0.02386	7.79319	(0.139;0.233)
$\hat{\lambda}_1^C$	-0.90822	0.02386	-38.0697	(-0.955;-0.861)
$\hat{\lambda}_1^C$	0.90822	0.02386	38.0697	(0.861;0.955)
$\hat{\lambda}_{11}^{AB}$	0.97248	0.02386	40.7632	(0.926;1.019)
$\hat{\lambda}_{21}^{AB}$	-0.97248	0.02386	-40.7632	(-1.019;-0.926)
$\hat{\lambda}_{11}^{AB}$	-0.97248	0.02386	-40.7632	(-1.019;-0.926)
$\hat{\lambda}_{21}^{AB}$	0.97248	0.02386	40.7632	(0.926;1.019)
$\hat{\lambda}_{112}^{C}$	-0.16635	0.02386	-6.97276	(-0.213;-0.119)
$\hat{\lambda}_{11}^{BC}$	0.16635	0.02386	6.97276	(0.119;0.213)
$\hat{\lambda}_{11}^{BC}$	0.16635	0.02386	6.97276	(0.119;0.213)
$\hat{\lambda}_{21}^{BC}$	-0.16635	0.02386	-6.97276	(-0.213;-0.119)
$\hat{\lambda}_{11}^{AC}$	0.05803	0.02386	2.43251	(0.011;0.105)
$\hat{\lambda}_{11}^{AC}$	-0.05803	0.02386	-2.43251	(-0.105;-0.011)
$\hat{\lambda}_{11}^{AC}$	-0.05803	0.02386	-2.43251	(-0.105;-0.011)
$\hat{\lambda}_{11}^{AC}$	0.05803	0.02386	2.43251	(0.011;0.105)
$\hat{\lambda}_{111}^{ABC}$	-0.12070	0.02386	-5.05917	(-0.167;-0.074)
$\hat{\lambda}_{112}^{ABC}$	0.12070	0.02386	5.05917	(0.074;0.167)
$\hat{\lambda}_{121}^{ABC}$	0.12070	0.02386	5.05917	(0.074;0.167)
$\hat{\lambda}_{122}^{ABC}$	-0.12070	0.02386	-5.05917	(-0.167;-0.074)
$\hat{\lambda}_{211}^{ABC}$	0.12070	0.02386	5.05917	(0.074;0.167)
$\hat{\lambda}_{212}^{ABC}$	-0.12070	0.02386	-5.05917	(-0.167;-0.074)
$\hat{\lambda}_{221}^{ABC}$	-0.12070	0.02386	-5.05917	(-0.167;-0.074)
$\hat{\lambda}_{222}^{ABC}$	0.12070	0.02386	5.05917	(0.074;0.167)

A = 1 ... age < 24 months
A = 2 ... age >= 24 months
B = 1 ... still breastfeeding
B = 2 ... not breastfeeding
C = 1 ... malnourished
C = 2 ... not malnourished

From Table 7.12 the negative coefficient $\hat{\lambda}_{111}^{ABC}$ (which represents children under 24 months that are still breastfeeding and are malnourished) indicates that there is a lower frequency count than expected. This result is expected since the effect of breastfeeding should decrease the number of malnourished children under 24 months. Similarly the positive coefficient $\hat{\lambda}_{112}^{ABC}$ indicates that there is a higher than expected count for children under 24 months that are not malnourished and still breastfeeding. Of interest is that breastfeeding works in the opposite direction above 24 months. This is reflected by the positive coefficient $\hat{\lambda}_{211}^{ABC}$ which indicates that there is a higher than expected count for children over 24 months that are malnourished and still breastfed. Children who are still breastfeeding at an older age might come from poorer households where other nutritional supplements are not affordable. The breastmilk provided might not be sufficient or nutritious enough for the older child and therefore these children tend to have more malnutrition. A positive $\hat{\lambda}_{222}^{ABC}$ further supports this hypothesis, since the number of children that are no longer breastfeeding over 24 months and are not malnourished is greater than expected. The remaining three way interaction terms follow a similar trend to those discussed above.

The Z-value and confidence intervals given in Table 7.12 need to be interpreted with caution owing to the restrictions on the coefficients i.e. coefficients for the different levels of a factor must sum to zero (from equation 7.4.3).

In this Chapter various risk factors such as gender and place of residence have a significant association with the three indicators of malnutrition at different ages. These results therefore do not support the proposal made by Dibley et al. (1987) (see Chapter 5), who called for separate analyses for below and above 24 months for anthropometric data.

In Chapter 8 a logistic regression model under complex sampling is fitted to the child nutrition data to predict malnutrition.

Besides age and breastfeeding other predictor variables such as source of income, place of residence, round, sex etc. are considered to find the best fitting model. Of interest is whether malnutrition relates to breastfeeding and age in a similar fashion to the log-linear model fitted above and whether other variables can be considered to be strong predictors of malnutrition.

CHAPTER 8

LOGISTIC REGRESSION IN THE CASE OF COMPLEX SURVEYS

8.1 INTRODUCTION

In the above explanation of the log-linear model, it can be seen that all the variables are considered to be explanatory. i.e. there is no distinction between a dependent variable and an independent variable. When there is a distinction to be made the usual linear regression model can be used where the dependent variable is continuous. However it is often the case that the dependent or outcome variable is dichotomous. Nutritional status is an example of a dichotomous variable, where a child is considered to be underweight if their weight-for-age Z-score is below -2 Z-scores, and normal if their weight-for-age Z-score is greater than or equal to -2 Z-scores (see Chapter 5). One might be interested in determining underweight children from risk factors such as breastfeeding or source of income. In this case a logistic regression model is applicable.

The computer package STATA (Stata Corporation(1993)) was initially used in this research for analysing the data of the NH&NS using a logistic regression model. STATA takes clustering (using Huber's formula, Huber (1967)) into consideration but not stratification. It is therefore implicitly assumed that the regressions are homogeneous across the strata. According to Deaton(1994), when there is reason to believe that the regressions are heterogenous across strata , "it is good practice to estimate separate regressions for each stratum, or at least for groups of strata". Deaton(1994) comments further that "When there are many strata, running separate regressions is at best clumsy; it generates too many numbers, and the results are hard to interpret, so that it makes sense to assume homogeneity". In the case of the NH&NS, where stratification is by place of residence, zone, and district, one would expect heterogeneity across strata at the place of residence (rural/urban) stratification level. Therefore two logistic regression models were fitted, one for the rural stratum, and one for the urban stratum of the NH&NS, in order to determine a model for predicting malnutrition for children under 5 years.

I acquired the SUDAAN (Sudaan (1995)) computer package shortly after analysing the data with STATA. This package has the advantage in that it caters for clustering as well as for stratification. Therefore all the results reported in this thesis using the logistic regression model are generated using SUDAAN. However, should one use a program that does not provide a facility for stratification, the approach described in the above paragraph can be applied.

8.2 COMPARING THE LOGISTIC REGRESSION MODEL WITH THE LINEAR REGRESSION MODEL

In the case of simple linear regression (i.e. where there is only one independent variable) the conditional expectation $E(y|x)$ is given by:

$$E(Y|x) = \beta_0 + \beta_1 x \qquad (8.2.1)$$

which implies that it is possible for $E(Y|x)$ to take on any value as x ranges between $-\infty$ and $+\infty$. With a dichotomous dependent variable Y, taking on the values of zero or 1, the conditional expectation

$$\begin{aligned} E(Y|x) &= 1.P[Y=1|x] + 0.P[Y=0|x] \\ &= P[Y=1|x] \qquad (8.2.2) \\ &= \pi(x) \text{ , say} \end{aligned}$$

must lie between zero and one (see Hosmer and Lemeshow(1989)).

The logit transformation,

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \qquad (8.2.3)$$

transforms the range of $\pi(x)$ from $(0,1)$ to $(-\infty,\infty)$ and is central to the logistic regression model. The linear logistic model assumes a linear model for $g(x)$ i.e.

$$g(x) = \beta_0 + \beta_1 x$$

which is equivalent to the nonlinear model:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \qquad (8.2.4)$$

for the conditional expectation.

In the usual linear regression model the assumption is made that a realization of the outcome variable may be expressed as $y = E(Y|x) + \epsilon$, where the error ϵ, which expresses an observation's deviation from the conditional mean, follows a normal distribution with mean zero and some variance that is constant across all levels of the independent variable. This is not the case in the logistic regression model defined as $y = \pi(x) + \epsilon$ where ϵ can assume only one of two possible values. When y=1 (with probability $\pi(x)$) then $\epsilon = 1 - \pi(x)$, and when y=0 (with probability $1 - \pi(x)$) then $\epsilon = -\pi(x)$, and has a binomial distribution with expectation

$E[\epsilon] = (1-\pi(x))x\pi(x) + (-\pi(x))x(1-\pi(x)) = 0$ and
variance $Var[\epsilon] = \pi(x)(1-\pi(x))$.

8.3 FITTING THE LOGISTIC REGRESSION MODEL

There are two steps in fitting the logistic regression model under cluster sampling:

i) Solutions are found to the weighted maximum likelihood equations which are identical to those obtained from a strictly binomial model. The standard binomial-based estimates of β are asymptotically normally distributed and consistent, even under cluster sampling. In other words, the regression coefficient estimates are robust to violations of model assumptions (Binder, 1983).

ii) The standard asymptotic variance estimates (obtained from the inverse of the information matrix) are however not robust to violations of model assumptions. Therefore Taylor linearization is applied for implicitly-defined parameter vectors, in conjunction with a between-cluster variance estimation formula (Binder, 1983).

Let

$y_{ijk} = 1$ if sample member ijk has the attribute of interest

$y_{ijk} = 0$ otherwise

where

$i = 1,2,\ldots m$ strata; $j = 1,2,\ldots,p$ PSU's or clusters; $k=1,\ldots,n_{ij}$ observations.

$\mathbf{x}_{ijk} = (1, x_{1,ijk}, \ldots, x_{q,ijk})'$ = vector of regression variables (stratum, psu, and observation specific) for the kth element within the cluster.

$\beta = (\beta_0, \beta_1, \ldots, \beta_q)'$ is a vector of unknown regression coefficients.

w_{ijk} = sampling weight for sample member ijk.

The Logistic Regression model is stated as follows:

$$P_{ijk}(\beta) = P(y_{ijk} = 1|\boldsymbol{x}_{ijk}, \beta) = [1 + \exp(-\boldsymbol{x}_{ijk}'\beta)]^{-1} = p_{ijk} \qquad (8.3.1)$$

$$\Rightarrow \frac{p_{ijk}}{1-p_{ijk}} = \exp(\boldsymbol{x}_{ijk}'\beta)$$

$$\Rightarrow \log\frac{p_{ijk}}{1-p_{ijk}} = \boldsymbol{x}_{ijk}'\beta \qquad (8.3.2)$$

The likelihood function expresses the probability of the observed data as a function of the unknown parameters. Thus for those pairs where $y_{ijk}=1$ the contribution to the likelihood function is $p(\mathbf{x}_{ijk})$, and for those pairs where $y_{ijk}=0$ the contribution to the likelihood function is $1-p(\mathbf{x}_{ijk})$. Even though the observations within a cluster are not independent, based on the arguments in (i) above, the obsevations are treated as n_{ij} independent Bernoulli observations (from cluster j of stratum i) whose contribution to the weighted likelihood function for the pair $(\mathbf{x}_{ijk}, y_{ijk})$ is through the term

$$\zeta(\boldsymbol{x}_{ijk}) = p_{ijk}{}^{w_{ijk}y_{ijk}}[1 - p_{ijk}]^{w_{ijk}(1-y_{ijk})}$$

$$= (\frac{p_{ijk}}{1-p_{ijk}})^{w_{ijk}y_{ijk}}(1-p_{ijk})^{w_{ijk}}$$

$$= \exp(\boldsymbol{x}_{ijk}'\beta w_{ijk}y_{ijk})[\exp(-\boldsymbol{x}_{ijk}'\beta)(1 + \exp(-\boldsymbol{x}_{ijk}'\beta))^{-1}]^{w_{ijk}}$$

$$(8.3.3)$$

and

$s_{ijk} = \log(1+\exp(-\mathbf{x}'_{ijk}\beta) = \log(r_{ijk})$

Therefore $ds_{ijk}/d\beta = dr_{ijk}/d\beta \,.\, 1/r_{ijk}$

where

$dr_{ijk}/d\beta = -\mathbf{x}_{ijk}\exp(-\mathbf{x}'_{ijk}\beta)$

and $1/r_{ijk} = 1/(1+\exp(-\mathbf{x}'_{ijk}\beta)$

Therefore $ds_{ijk}/d\beta = -\mathbf{x}_{ijk}\exp(-\mathbf{x}'_{ijk}\beta)(1+\exp(-\mathbf{x}'_{ijk}\beta))^{-1}$

$$= -\mathbf{x}_{ijk}(1-p_{ijk}) \qquad (8.3.6)$$

We need to solve

$$W(\beta) = \frac{\partial l(\beta)}{\partial(\beta)} = 0$$

$$i.e.\ \sum_i\sum_j\sum_k \{w_{ijk}\mathbf{x}_{ijk}y_{ijk} - w_{ijk}\mathbf{x}_{ijk} - w_{ijk}[-\mathbf{x}_{ijk}(1 - p_{ijk})]\}$$

$$= \sum_i\sum_j\sum_k w_{ijk}\mathbf{x}_{ijk}y_{ijk} - \sum_i\sum_j\sum_k w_{ijk}\mathbf{x}_{ijk}p_{ijk} = 0 \qquad (8.3.7)$$

This system of implicit nonlinear equations in the q+1 unknowns,

$$\underline{\beta} = (\beta_0\ \beta_1\ \ldots\ \beta_q)$$

is usually solved by a conditional Newton-Raphson procedure that determines the value $\hat{\beta}$ of β that maximizes $l(\hat{\beta})$ based on the iterative corrections:

$$\beta^{(s+1)} = \beta^{(s)} - (H^{(s)})^{-1}q^{(s)} \qquad (8.3.8)$$

where

$$q^{(s)} = \frac{\partial l(\beta)}{\partial \beta\big|_{\beta^{(s)}}} = \left(\frac{\partial l(\beta)}{\partial \beta_0\big|_{\beta^{(s)}}} \cdots \frac{\partial l(\beta)}{\partial \beta_q\big|_{\beta^{(s)}}}\right)'$$

$$= \sum_i \sum_j \sum_k w_{ijk}\boldsymbol{x}_{ijk}y_{ijk} - \sum_i \sum_j \sum_k w_{ijk}\boldsymbol{x}_{ijk}p_{ijk}^{(s)} \qquad (from(8.3.7))$$

$$(8.3.9)$$

where

$$p_{ijk}^{(s)} = [1 + \exp(-\boldsymbol{x}_{ijk}\beta_{(s)})]^{-1}$$

and

$$H^{(s)} = \frac{\partial^2 l(\beta)}{\partial l(\beta)\partial l(\beta)'\big|_{\beta^{(s)}}}$$

$$= \frac{\partial(\sum_i \sum_j \sum_k w_{ijk}\boldsymbol{x}_{ijk}y_{ijk} - \sum_i \sum_j \sum_k w_{ijk}\boldsymbol{x}_{ijk}p_{ijk}^{(s)})}{\partial \beta}$$

$$= \sum_i \sum_j \sum_k -\boldsymbol{x}_{ijk}\boldsymbol{x}_{ijk}'w_{ijk}\exp(-\boldsymbol{x}_{ijk}'\beta^{(s)})(1 + \exp(-\boldsymbol{x}_{ijk}'\beta^{(s)})^{-2}$$

$$= \sum_i \sum_j \sum_k -\boldsymbol{x}_{ijk}\boldsymbol{x}_{ijk}{}' w_{ijk} d_{ijk}^{(s)} \quad (\mathit{from}\ (8.3.4)) \qquad (8.3.10)$$

are the terms evaluated at $\beta^{(s)}$, the sth guess for $\hat{\beta}$ (Agresti 1990),

where

$$d_{ijk}^{(s)} = p_{ijk}^{(s)}(1 - p_{ijk}^{(s)})$$

Substituting (8.3.9) and (8.3.10) into (8.3.8) yields the system of equations for the iterative solution for the maximum likelihood estimator of β:

$$\beta^{(s+1)} = \beta^{(s)} + [\sum_i \sum_j \sum_k w_{ijk}\boldsymbol{x}_{ijk}\boldsymbol{x}'_{ijk} d_{ijk}^{(s)}]^{-1} [\sum_i \sum_j \sum_k w_{ijk}\boldsymbol{x}_{ijk}(y_{ijk} - p_{ijk}^{(s)})]$$

(8.3.11)

The default starting value vector $\hat{\beta}_0$ suggested in Sudaan(1995) for the first iteration (s=1) is the null vector of length (q+1).

8.4 VARIANCE ESTIMATION UNDER CLUSTER SAMPLING

Under cluster sampling with nonzero intracluster correlation, the standard asymptotic covariance matrix estimate given by J^{-1} (the inverse of the weighted sample information matrix) which are derived in the same way as in (8.3.10), where

$$J = -H = -[\frac{\partial^2 l(\beta)}{\partial \beta^2}] = \sum_i \sum_j \sum_k \boldsymbol{x}'_{ijk}\boldsymbol{x}_{ijk} w_{ijk} \hat{d}_{ijk} \qquad (8.4.1)$$

are biased estimates (Binder(1981)).

Instead, an implicit differentiation method for estimating the covariance matrix for a vector of survey statistics was proposed and justified by Binder (1981, 1983) (see Appendix H).

Binder(1981,1983) uses the Taylor linearization for approximating the variance of implicitly defined parameter vectors:

$$\hat{var}(\hat{\beta}) = (J^{-1})\, \hat{var}[\hat{W}(\hat{\beta})]\, (J^{-1})' \qquad (8.4.2)$$

where

$$\hat{W}(\hat{\beta}) = \frac{\partial l(\beta)}{\partial\beta}\Big|_{\beta=\beta} = \sum_{i}\sum_{j}\sum_{k} \hat{W}(z_{ijk};\beta) \qquad (8.4.3)$$

is the score function which is a simple linear function of the observations and the linearized variate vector (see equation 8.3.7). For the ijk-th unit this function is:

$$z_{ijk} = \mathbf{x}_{ijk}(y_{ijk}-\hat{p}_{ijk}) \qquad (8.4.4)$$

To compute the between-PSU within-stratum variance estimate for a vector of linear statistics, accumulations of the linearized variate vectors are first formed at the PSU level

$$z_{ij} = \sum_{k} z_{ijk}, \qquad k = 1,\ldots,n_{ij}$$

Then the between-PSU within stratum mean square deviations matrix is formed by:

$$S_z = \sum_i m_i S_{zi},$$

where m_i = the number of PSU's in stratum i, and the sample mean squares and cross-products matrix S_{zi} is given by:

$$S_{zi} = \frac{\sum_{j=1}^{m_i} (z_{ij} - \overline{z_i})(z_{ij} - \overline{z_i})'}{(m_i - 1)} \qquad (8.4.5)$$

$$\overline{z_i} = \sum_j \frac{z_{ij}}{m_i}$$

From Appendix H the estimated cluster covariance matrix for $\hat{\beta}$ is then given by:

$$v\hat{a}r(\hat{\beta}) = (J^{-1}) S_z (J^{-1})' \qquad (8.4.6)$$

8.5 HYPOTHESIS TESTING

The null hypothesis for testing whether the regression coefficient associated with the ith variable is zero, is:

$H_0 : \beta_i = 0$ vs. $H_A : \beta_i \neq 0$

The overall null hypothesis is:

$H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$

One can also specify linear combinations of the model parameters for testing i.e.

$H_0 : C\beta = 0$

where C is a contrast matrix. See Chapter 7 for a more detailed discussion of the tests that can be applied to the above hypotheses.

The Wald test statistic for testing the third null hypothesis is:

$$\chi^2 = [C\beta]'[C\hat{Var}(\beta)C']^{-1}[C\beta]$$

which under H_0 has a chi-squared distribution with c degrees of freedom, where c = rank of C.

8.6 INTERPRETATION OF THE COEFFICIENTS

In order to describe the interpretation of the logistic regression coefficients, the simplest scenario is considered where the independent variable is dichotomous. This scenario forms a "conceptual foundation" for all other situations (see Hosmer and Lemeshow(1989)). Consider the 2x2 table in Table 8.1, where the independent variable x is coded as either zero or 1.

Table 8.1 **Values of the Logistic Regression Model when the Independent Variable is Dichotomous**

	$x = 1$	$x = 0$
$y = 1$	$\pi(1)$	$\pi(0)$
$y = 0$	$1 - \pi(1)$	$1 - \pi(0)$

where

$$\pi(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$$

$$\text{and } \pi(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \qquad (8.6.1)$$

The odds of the outcome being present among individuals with x=1 is defined as $\pi(1)/[1 - \pi(1)]$. Similarly, the odds of the outcome being present among individuals with x=0 is defined as $\pi(0)/[1 - \pi(0)]$. The log of the odds, or logits, are

$$g(1) = \ln\{\pi(1)/[1 - \pi(1)]\} = \beta_0 + \beta_1$$

$$\text{and} \quad g(0) = \ln\{\pi(0)/[1 - \pi(0)]\} = \beta_0 \qquad (8.6.2)$$

The odds ratio ψ is defined as the ratio of the odds for x=1 to the odds for x=0, and is given by the equation

$$\psi = (\pi(1)/[1 - \pi(1)]) \;/\; (\pi(0)/[1 - \pi(0)])$$

$$= \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1} \qquad (8.6.3)$$

and the log-odds ratio is given by

$$\ln \psi = \ln\{(\pi(1)/[1 - \pi(1)]) \;/\; (\pi(0)/[1 - \pi(0)])\}$$

$$= g(1) - g(0)$$

$$= \beta_1 \qquad (8.6.4)$$

which is the logit difference.

For example if y denotes the presence or absence of malnutrition in a child under two years old, and if x denotes whether or not the child is still being breastfed, then the log-odds ratio for the logit

$g(x) = 0.5 + 0.7x$

is given by $\beta_1=0.7$ with an odds ratio of $\psi = \exp(0.7) = 2.014$. This indicates approximately that malnutrition occurs twice (2.014 times) as often among children that are 'not still breastfed' than among children that are 'still breastfed'.

For purposes of developing the method to interpret the coefficient for a continuous variable, the assumption is made that the logit is linear in the continuous predictor variable, x. The equation for the logit is $g(x) = \beta_0 + \beta_1 x$. It follows that the slope coefficient, β_1, gives the change in the log odds for an increase of a unit in x, i.e. $\beta_1 = g(x+1) - g(x)$ for any value of x (Hosmer and Lemeshow 1989). The value of a unit is not always meaningful. For example if the range of x is from zero to one, then a change of 1 is too large and a change of 0.01 might be more reasonable. Therefore the log odds for a change of c units in x is obtained from the logit difference $g(x+c) - g(x) = c\beta_1$ and the associated odds ratio is obtained by exponentiating this logit difference, $\psi(c) = \psi(x+c,x) = \exp(c\beta_1)$. Therefore in summary, an estimated log odds ratio is also used in the interpretation of the estimated coefficient for a continuous variable as is the case for dichotomous variables (see equation (8.6.4)). The only difference being that a meaningful change must be defined for continuous variables.

In the case of independent variables that are discrete and nominally ordered such as 'child minder' in the NH&NS (see Appendix F), the method of choice is to use a set of design variables. The two methods to create design variables are the partial method and the marginal method. The partial method is the most commonly employed method appearing in the literature (Hosmer and Lemeshow, 1989). The primary reason for the use of this method is the interest in estimating the risk of an "exposed" group relative to that of a "control" or "unexposed" group. The partial method or reference cell parametrization can be explained using the following example. The variable 'child minder' can take on 7 different values, so that six design variables are required. For instance if the response is "mother" then all 6 of the design variables take on the value zero. The design variables for the variable 'child minder' can be seen in Table 8.2. Exponentiation of the estimated coefficient for the design variable D1 would therefore estimate the odds ratio of 'grandmother' relative to 'mother'.

Table 8.2 Design variables created from the variable 'child minder' using the partial method

	Design Variable					
CHILDMIND	D1	D2	D3	D4	D5	D6
mother	0	0	0	0	0	0
grandma	1	0	0	0	0	0
aunt	0	1	0	0	0	0
maid	0	0	1	0	0	0
father	0	0	0	1	0	0
sister	0	0	0	0	1	0
other	0	0	0	0	0	1

The marginal method of coding design variables, which is more frequently used in analysis of variance and linear regression than in logistic regression is called "deviation from mean coding". This expresses effect as the deviation of the "group mean" from the "overall mean". In the case of logistic regression the "group mean" is the logit for the group and the "overall mean" is the average logit. The required coding is obtained as illustrated for childminder in Table 8.3, by setting the value of all the design variables equal to -1 for one of the categories, and then using the 0,1 coding for the remainder of the categories. In applying the marginal method "the interpretation of the estimated coefficients is not as easy or clear as in the situation when a referent group is used" (Hosmer and Lemeshow (1989)). Therefore based on the above arguments the partial method will be used in all further analyses in this thesis.

Table 8.3 Design variables created from the variable 'child minder' using the marginal method

	Design Variable					
CHILDMIND	D1	D2	D3	D4	D5	D6
mother	-1	-1	-1	-1	-1	-1
grandma	1	0	0	0	0	0
aunt	0	1	0	0	0	0
maid	0	0	1	0	0	0
father	0	0	0	1	0	0
sister	0	0	0	0	1	0
other	0	0	0	0	0	1

8.7 INCLUDING INTERACTIONS IN THE LOGISTIC REGRESSION MODEL

Consider a model containing a dichotomous risk factor F, a continuous covariate X and their interaction FX. Suppose F has the levels f_1 and f_0. The logit for this model at F=f and X=x is therefore:

$$g(f,x) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 fx \qquad (8.7.1)$$

Therefore the log-odds ratio for $F=f_1$ versus $F=f_0$ with X held constant at X=x is

$$\ln[\psi(F=f_1, F=f_0, X=x)] = g(f_1,x) - g(f_0,x) \\ = \beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0) \qquad (8.7.2)$$

Therefore in this case when F is a dichotomous risk factor with $f_0=0$ and $f_1=1$,

$$\ln[\psi(F=1, F=0, X=x)] = \beta_1 + \beta_3 x \qquad (8.7.3)$$

and the estimated variance of the estimated log odds ratio is

$$\hat{var}\ln[\hat{\psi}(F=1,F=0,X=x)] = \hat{var}(\hat{\beta}_1) + \hat{var}(\beta_3)x^2 + 2\hat{cov}(\beta_1,\beta_3)x \qquad (8.7.4)$$

Therefore to correctly assess the risk factor on the response variable, one must include the interaction of this risk factor with the covariate in the model. A similar example can be found in Section 8.9.

8.8 MEASURES OF GOODNESS-OF-FIT

8.8.1 R^2-TYPE MEASURES

The R^2 currently provided by the SUDAAN program is calculated as the weighted simple correlation between the observed and predicted values. The absolute value of this R-square value is irrelevant. It is only useful to compare different models with the same data. A planned enhancement to SUDAAN is to produce the same R^2 measure as that of STATA.

This R^2-type measure, referred in the STATA computer package as a Pseudo R^2 is given by

$$R^2_1 = 100(l_q - l_0)/(l_s - l_0) \qquad (8.8.1)$$

where l_0, l_q, and l_s (see Section 8.3) are respectively the log-likelihoods for the model containing only the intercept, the model containing the intercept plus q covariates, and for the saturated model. Probabilities are no greater than 1.0, so log likelihoods are nonpositive. As the model complexity increases, the parameter space expands, so the value of the maximized likelihood increases. This measure falls between 0 and 1. It equals 0 when the model provides no improvement in fit over the null model, and it equals 1 when the model fits as well as the saturated model. This measure is not advocated by Hosmer and Lemeshow(1989), who describe R^2_1 as "nothing more than an expression of the likelihood ratio test and, as such, is not a measure of goodness-of-fit".

Based on the above arguments for .ot using the R^2 provided by STATA and SUDAAN, for the purposes of this thesis therefore, the R^2 of SUDAAN will only be reported and commented on when comparing models that are fitted to the same data. The more appropriate measures of goodness-of-fit, the Pearson Chi-Square and the Hosmer and Lemeshow test, discussed in Sections 8.8.2 and 8.8.3, respectively, will be used instead. Since SUDAAN does not provide these tests, the predicted and observed observations from the SUDAAN analysis were transformed into a SAS data set and programs were written in SAS to perform the tests (see Appendix J for the Pearson chi-square goodness-of-fit test and Appendix K for the Hosmer Lemeshow test).

8.8.2 PEARSON CHI-SQUARE AND DEVIANCE

Let y_i denote the number of successes in n trials at the ith of I settings or combinations of the predictor variables. For a logistic regression model, standardised residuals for the fits provided by the I binomial distributions are

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{[n_i\hat{\pi}_i(1 - \hat{\pi}_i)]^{1/2}}, \quad i=1,\ldots,I \tag{8.8.2}$$

Each standardised residual e_i, compares the difference between an observed count and its expected value with the estimated standard deviation of the observed count (Agresti, 1990). The standardised residuals e_i are often treated like standard normal deviates, with absolute values larger than 2 indicating possible lack of fit.

The Pearson statistic for testing the fit of the model is computed from $\{e_i\}$ by

$$X^2 = \sum e_i^2 \tag{8.8.3}$$

The distribution of X^2 is chi-square with degrees of freedom equal to I-(p+1), where p is the number of predictor variables in the model.

This X^2 goodness-of-fit statistic does not have an approximate chi-squared distribution when applied to a logistic regression model with a continuous predictor variable, unless there are many observations at each observed level of the predictor variable. When this is not the case and there are not many observations at each observed level of the continuous predictor variable, a test by Hosmer and Lemeshow (1980) is more appropriate.

8.8.3 THE HOSMER-LEMESHOW TEST

The Hosmer-Lemeshow goodness-of-fit statistic $\hat{C}$, is given by

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

where n'_k is the number of predictor variable patterns in the kth out of g groups,

$$o_k = \sum_{j=1}^{n'_k} y_j$$

is the number of positive responses among the n'_k predictor variable patterns in group k, and

$$\bar{\pi}_k = \sum_{j=1}^{n_k} \frac{m_j \hat{\pi}_j}{n_k}$$

is the estimated average probability in group k (Hosmer and Lemeshow 1980), where m_j is the number of distinct combinations of the predictor variables such that

$$\sum_{j=1}^{n'_k} m_j = n'_k$$

The grouping is based on a percentile-type of grouping, usually with 10 groups. This results in the first group containing the $n'_1 = n/10$ subjects having the smallest estimated probabilities, and the last group containing the $n'_{10}=n/10$ subjects having the largest estimated probabilities (Hosmer and Lemeshow 1989).

The distribution of the $\hat{C}$ statistic is approximately a chi-square distribution with g-2 degrees of freedom.

Therefore in all subsequent analyses of this thesis, the Pearson chi-square will be applied to test the adequacy of the logistic regression model only if all the predictor variables are categorical. If there are one or more continuous predictor variables (and any number of discrete ones), then the Hosmer-Lemeshow test will be used.

8.9 ANALYSING CHILD NUTRITION STATUS USING A LOGISTIC REGRESSION MODEL

In order to demonstrate the model building process using the logistic regression model, consider again the child nutrition section of the NH&NS. A logistic regression model to predict the acutely malnourished proportion of the population could not be fitted to this data set. From Tables 7.3 to 7.7 of Chapter 7 it is clear that the acutely malnourished proportion ranges between 2% and 5% of the sample. There was not enough variation in the outcome category weight-for-height less than a Z-score of -2 (which represents the acutely malnourished) for a logistic regression model to be fitted to the data. In the following example a logistic regression model under complex sampling is therefore fitted to predict low weight-for-age (underweight children) instead.

A child is considered to be underweight if its weight-for-age Z-score is below -2, and normal otherwise. The variables given in Table 8.4 are used as possible predictors of an underweight child in the logistic regression model (see also Appendix E).

Since there are two distinctly different reference populations used for child nutrition data (Fels and NCHS) resulting in discontinuities at 24 months of age, and considering the recommendations given in Chapter 5 Section 5.5 (Dibley et al. 1987b), two models were fitted to the above data, one for children below 24 months, and one for children 24 months or older. Thereafter a single model for all children under the age of five years was also fitted, in order to arrive at the best approach to fitting a logistic regression model to such data.

SUDAAN does not have the programming facility to perform a stepwise logistic procedure, therefore a stepwise analysis (backward selection) was performed by re-running SUDAAN many times, identifying after each run the most non-significant predictor or interaction term and then re-running the model without that term in the model. Consider firstly the model for children under 24 months. Initially there are 10 main effects and 28 two-way interaction terms included in the model. After the first run the interaction effect sex X sfed, with the largest p-value for the Satterthwaite Adjusted Chi-square statistic (p=0.7001), was excluded ırom the model. The model was then refitted and the least significant term after the second run was the interaction term educ X age (p=0.6790). This term was then excluded from the model. This procedure continued until all the effects in the model were significant at the 5% level and the resultant model in Table 8.5 was obtained. It can be seen that with the exception of the childminder variable 'care', and 'place of residence' each of the predictor variables as well as the age x sfed interaction has a significant effect and are included in the model. It is surprising that being in a rural or urban district of Lesotho does not have a significant effect on the malnutrition rate of children. A possible reason for this effect not being significant will be investigated later in this Chapter.

Table 8.4 Description of the Variables used in the Logistic Regression Model

Variable	Description
sex=1	female
sex=2 reference category	male
wazz=1	malnourished (weight-for-age Z-score < -2)
wazz=0	not malnourished (weight-for-age Z-score >=-2)
educm=1	Mother below Standard 5 education (uneducated)
educm=2 reference category	Mother above Standard 5 education (educated)
educf=1	Father below Standard 5 education (uneducated)
educf=2 reference category	Father above Standard 5 education (educated)
rurban=1	rural
rurban=2 reference category	urban
sfed=1	Still breastfed=Yes
sfed=2 reference category	Still breastfed=No
age	age in months
source=1	income from subsistence farming
source=2	income of cash cropping or livestock sales
source=3	Business income
source=4	Wages or salaries in cash
source=5 reference category	Cash remittances from migrant workers
care=1	Childminder - mother
care=2	Childminder - Grandmother
care=3	Childminder - Aunt
care=4	Childminder - Maid
care=5	Childminder - Sister
care=6	Childminder - Father
care=7 reference category	Childminder - Other person
round=2	June round
round=3	November round
round=4 reference category	February round

Since the continuous predictor variable age was included in the model the Hosmer Lemeshow test was applied to test the goodness-of-fit of the model (see Section 8.8.3 and Appendix K). In this case $\hat{C}$=25.3998 > 15.507, the 95% critical value of the chi-square distribution with eight degrees of freedom (there are 10 groups, therefore g=10-2=8). This indicates that the model in Table 8.5 does not fit the data well and possibly additional predictor variables which were not considered in the backward selection method need to be included in the model. From Table 8.5 it can be seen that age is a highly significant predictor of malnutrition and a graph of the observed proportion malnourished in each age category (0-24 months) is plotted in Figure 8.1. The plot indicates that higher order polynomial terms in age may need to be included in the model. The large variation in the observed proportion of malnourished children of the non-breastfeeders is mainly due to the low numbers in these age categories.

Table 8.5 Tests of Main Effects and Interactions in the Multiple Logistic Regression Model for Weight-for-Age under 24 months

weight-for-age	Degrees of freedom	S_waite Adj Df	S_waite Adj Chisq	p-value S_waite Adj Chisq
overall model	10	7.738	917.89	0.000
model minus intercept	9	7.744	142.21	0.000
sfed	1	1.000	5.04	0.024
sex	1	1.000	34.21	0.000
age	1	1.000	44.15	0.000
educm	1	1.000	4.65	0.031
source	4	3.860	12.62	0.012
age X sfed	1	1.000	11.07	0.000

R^2=0.038826

$\hat{C}$=25.3988, df=8, p-value= 0.0013301

For the re-run of the model, the backward selection process was again applied, this time including polynomial terms up to the fifth power in age. The resultant model from this fitting process can be seen in Tables 8.6 and 8.7. In Table 8.6 it can be seen that an age^2 term and the interactions age^3 X sfed and age^4 X sfed are now also included in the model. For this model R^2=0.05432 is larger than R^2=0.038826 for the model in Table 8.5 which indicates a better fit. The value of the Hosmer-Lemeshow goodness-of-fit statistic is $\hat{C}$=9.14112, and the corresponding p-value computed from the chi-square distribution with 8 degrees of freedom is 0.33053. This suggests that the model fits the data quite well.

FIGURE 8.1 Plot of the observed proportion malnourished in each age category for age < 24 month

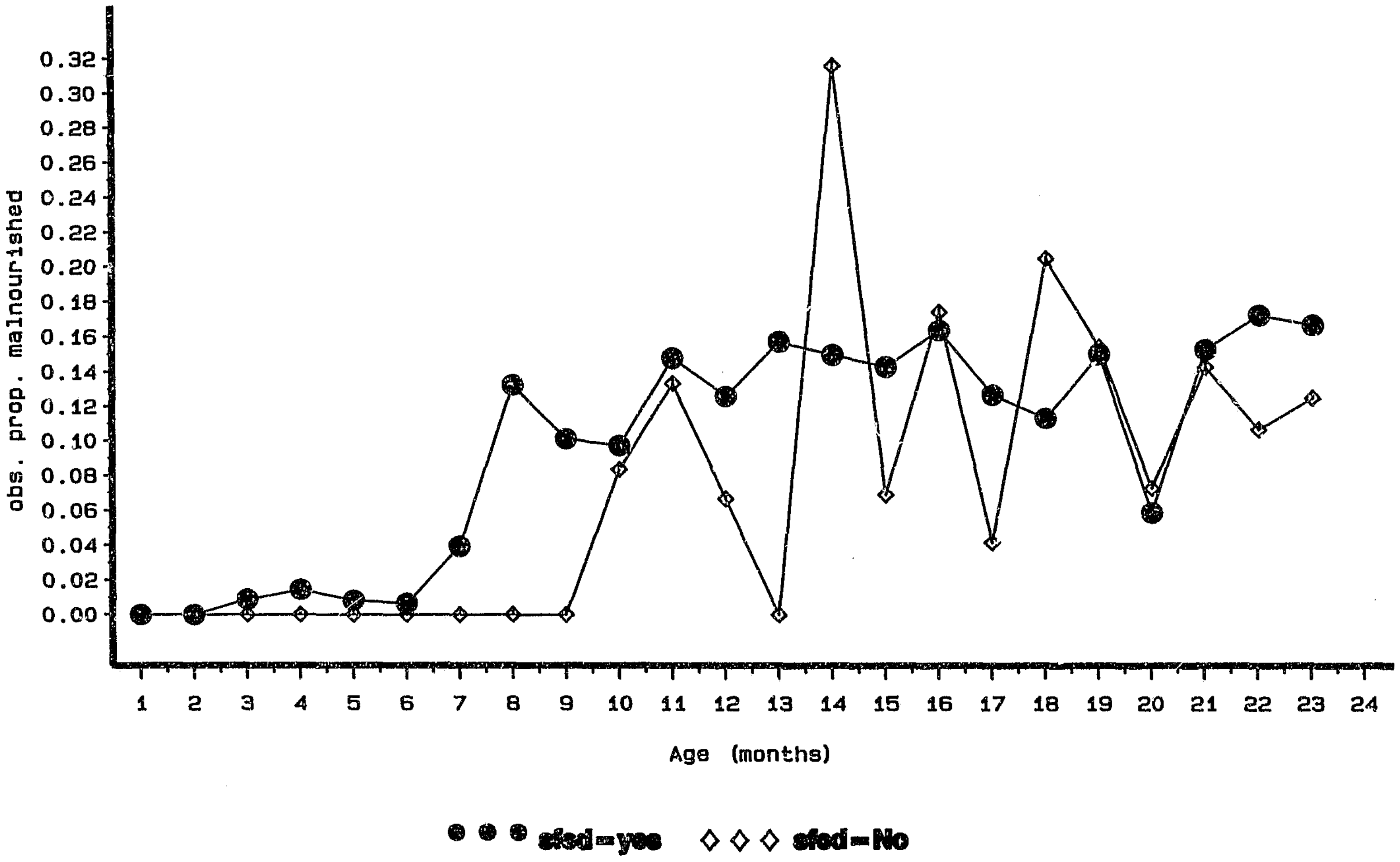


Table 8.6 Tests of Main Effects and Interactions in the Multiple Logistic Regression Model for Weight-for-Age under 24 months, including higher order polynomial terms in age

weight for age	Degrees of freedom	S_waite Adj DF	S_waite Adj Chisq	p-value S_waite Adj Chisq
overall model	12	8.65	718.21	0.000
model minus intercept	11	9.00	107.70	0.000
sfed	1	1.00	2.370	0.124
age	1	1.00	60.388	0.000
age^2	1	1.00	32.189	0.000
sex	1	1.00	33.632	0.000
educm	1	1.00	4.839	0.028
source	4	3.85	13.212	0.009
age^3 X sfed	1	1.00	7.926	0.005
age^4 X sfed	1	1.00	8.756	0.003

The reference cell parametrization (see Section 8.6) has been used to fit the model, with the last level of each categorical predictor forming the reference cell. These coefficients are set to zero, but are retained on the beta vector. Table 8.7 provides the vector of estimated regression coefficients, their estimated standard errors, t-tests and p-values for testing whether each individual regression coefficient is significantly different from zero, estimated odds ratios and their 95% confidence limits.

Table 8.7 Estimated Coefficients, Estimated Standard Errors, T-statistics, p-values, Odds Ratios and their confidence intervals for the Multiple Logistic Regression Model containing the significant predictor variables and the two significant interactions identified by the Backward selection analyses - children under 24 months

weight for age	Beta Coeff	SE Beta	T-test Beta=0	P-value T-test Beta=0	Odds Ratio	Odds Ratio 95% CI
constant	-6.497	0.668	-9.720	0.000	0.00151	(0.00040;0.00570)
sfed						
1	-0.678	0.441	-1.54	0.128	0.50742	(0.21116;1.21937)
0	0.000	0.000	.	.	1.00000	(1.00000;1.00000)
age	0.743	0.096	7.77	0.000	2.10266	(1.73832;2.54337)
age^2	-0.022	0.004	-5.67	0.000	0.97843	(0.97098;0.98594)
sex						
1	-0.688	0.119	-5.80	0.000	0.50252	(0.39685;0.63633)
0	0.000	0.000	.	.	1.00000	(1.00000;1.00000)
educm						
1	0.389	0.177	2.20	0.031	1.47537	(1.03787;2.09729)
0	0.000	0.000	.	.	1.00000	(1.00000;1.00000)
source						
1	0.349	0.142	2.45	0.016	1.41761	(1.06872;1.88039)
2	0.869	0.232	3.74	0.000	2.38461	(1.50345;3.78220)
3	0.237	0.392	0.60	0.547	1.26769	(0.58086;2.76666)
4	0.283	0.235	1.20	0.233	1.32730	(0.83092;2.12021)
5	0.000	0.000	.	.	1.00000	(1.00000;1.00000)
age^3 X sfed						
sfed=1	$-7x10^{-4}$	$2x10^{-4}$	-2.81	0.006	0.99929	(0.99878;0.99979)
sfed=0	0.000	0.000	.	.	1.00000	(1.00000;1.00000)
age^4 X sfed						
sfed=1	$2x10^{-5}$	$1x10^{-5}$	2.96	0.004	1.00003	(1.00001;1.00005)
sfed=0	0.000	0.000	.	.	1.00000	(1.00000;1.00000)

R^2=0.05432

$\hat{C}$=9.14112, df=8, p-value=0.33053

Even though the age^3 X sfed and the age^4 X sfed interactions are included in the model in Table 8.7, contrary to the normal practice of including in the model the main effects constituting an interaction term, the main effects age^3 and age^4 are not included in the model. The reason for this exclusion is that once these main effects were included in the model, the interaction terms were no longer significant and neither were the main effects age^3 and age^4. The best fitting model where all the terms are significant can be seen in Table 8.7.

From Table 8.7 the odds of malnutrition for females versus males is 0.50252. In other words malnutrition occurs half as often among females than among males in the study population.

A lower education of the mother increases the odds of malnutrition. The education level of the father was not significant and therefore not included in the model. This result is as expected since in most cases children live with their mother and not necessarily with their father (from the NH&NS, of the adults who were absent from their household for two weeks prior to the survey, 73.21% were male and only 26.79% were female) and since a mother generally takes care of her child's health and nutrition on a daily basis. The overall test for the variable 'source of income' is significant. Relative to 'Cash remittance from migrant workers' (source5), the significant effects are 'Subsistence farming' (Source1: Odds Ratio=1.417607) and 'Sales of Livestock' (Source2: Odds Ratio=2.384608) which are associated with increases in malnutrition. This result suggests that households relying on the income from migrant workers have a higher living standard and therefore less child malnutrition than households where the source of income is from subsistence farming or the sales of livestock. The combined effect of the age^3 X sfed and age^4 X sfed interaction terms are observed by substituting all possible values for the predictor variables in Table 8.7. A maximum is reached at 15 months for breastfeeders with a probability of being malnourished = 0.352, and at 23 months for non-breastfeeders with a probability of being malnourished = 0.318. A plot of this model can be seen later in this Chapter in Figure 8.4 in which the effect of the interaction terms are displayed more clearly.

The same procedure as above was followed to fit the logistic regression model for children aged 24 months and above and the results can be seen in Tables 8.8 and 8.9. Only four main effects are found to be significant and therefore included in the model, namely: breastfeeding, education, round and source of income (there were no significant interaction terms). Since there ar no continuous predictor variables the more appropriate goodness-of-fit test is the Pearson chi-square test (see Section 8.8.2 and Appendix J). For this test X^2=64.5504 on 55 degrees of freedom (the number of combinations of the predictor variables minus the number of parameters in the model minus 1). The corresponding p-value is 0.15411, which suggests that the model is adequate.

Table 8.8 Tests of Main Effects and interaction in the Multiple Logistic Regression Model for Weight-for-Age of children over 24 months

weight for age	Degrees of freedom	S_waite Adj DF	S_waite Adj Chisq	p-value S_waite Adj Chisq
overall model	9	7.25	1151.03	0.000
model minus intercept	8	6.55	43.80	0.000
sfed	1	1.00	23.87	0.000
educm	1	1.00	6.08	0.014
round	2	1.91	9.22	0.009
source	4	3.75	18.94	0.001

A clearly different model from the under 24 month model is fitted for children over 24 months. The variables age, the higher order polynomials in age and sex are not significant nor are any interaction terms, round is significant and breastfeeding has the opposite effect from that in children under 24 months. Relative to 'Cash remittance from migrant workers' the only significant effect is 'Subsistence farming' (Source 1) which, as in the under 24 month model of Table 8.7, has an increasing effect on malnutrition.

Table 8.9 Estimated Coefficients, Estimated Standard Errors, T-statistics, p-values, Odds Ratios and their confidence intervals for the Multiple Logistic Regression Model containing the significant variables for the Backward selection analyses of children over 24 months

weight for age	Beta Coeff	SE Beta	T-test Beta=0	P-value T-test Beta=0	Odds Ratio	Odds Ratio 95% CI
constant	-2.466	0.177	-13.9	0.000	0.085	(0.060;0.121)
sfed						
1	0.611	0.125	4.89	0.000	1.842	(1.436;2.361)
2	0.000	0.000	.	.	1.000	(1.000;1.000)
educm						
1	0.304	0.123	2.47	0.016	1.355	(1.060;1.732)
2	0.000	0.000	.	.	1.000	(1.000;1.000)
round						
2	0.243	0.108	2.24	0.028	1.275	(1.027;1.582)
3	0.369	0.121	3.06	0.003	1.447	(1.138;1.840)
4	0.000	0.000	.	.	1.000	(1.000;1.000)
source						
1	0.399	0.104	3.84	0.000	1.491	(1.212;1.833)
2	0.363	0.235	1.54	0.127	1.437	(0.900;2.296)
3	-0.394	0.286	-1.38	0.173	0.674	(0.381;1.192)
4	0.168	0.140	1.20	0.233	1.183	(0.896;1.561)
5	0.00	0.000	.	.	1.000	(1.000;1.000)

R^2 = 0.0148, X^2=64.5504 , df=54, p-value=0.15411

Relative to February (round 4) malnutrition is 1.275 more times as likely to occur in June (round 2) and 1.447 times more likely to occur in November (round 3). This trend in the malnutrition rate is not what is expected according to Huss-Ashmore and Goodman(1989), who suggest that the rainy season of February should have a higher proportion of malnourished children than the colder harvesting season of June and the spring season of October and November. An odds ratio of 1.84 for sfed indicates that breastfeeding for this group is associated with increased malnutrition.

The results from the two fitted models shown above (Tables 8.7 and 8.9) clearly supports the associations found in Chapter 7 between malnutrition, breastfeeding, and age. The results further support the hypothesis that children who are still breastfeeding over the age of 24 months have a higher chance of being malnourished than those that are not breastfeeding.

A single model for all children under five years has also been fitted to the data in Table 8.10. In order to accommodate the two clearly different models for children below and above 24 months, the single model requires polynomial terms up to the fourth power in age as well as their interaction with breastfeeding to explain the two trends in malnutrition shown in Tables 8.7 and 8.9. This model fits the data well and there is not a significant lack-of-fit ($\hat{C}$=7.9427, p-value=0.4391).

Table 8.10 **Estimated Coefficients, Estimated Standard Errors, T-tests, and Odds Ratios for the Multiple Logistic Regression Model for all children under five years.**

weight for age	Beta Coeff	SE Beta	T-test Beta=0	P-value T-test Beta=0	Odds Ratio	Odds Ratio 95% CI
constant	-4.038159	1.018683	-3.96	0.000	0.017630	(0.002323;0.133807)
sfed						
1	-2.722320	1.022653	-2.66	0.009	0.065722	(0.008591;0.502771)
0	0.000000	0.000000	.	.	1.000000	(1.000000;1.000000)
age	1.348257	0.217880	6.18	0.000	3.850707	(2.496169;5.940280)
age^2	-0.077343	0.014732	-5.25	0.000	0.925572	(0.898837;0.953102)
age^3	0.001865	0.000395	4.73	0.000	1.001866	(1.000000;1.000000)
age^4	-0.000015	0.000004	-4.27	0.000	0.999985	(0.999978;0.999992)
sex						
1	-0.828752	0.161492	-5.13	0.000	0.436594	(0.316617;0.602033)
0	0.000000	0.000000	.	.	1.000000	(1.000000;1.000000)
educm						
1	0.347920	0.096598	3.60	0.000	1.416120	(1.168506;1.716203)
0	0.000000	0.000000	.	.	1.000000	(1.000000;1.000000)
round						
2	0.273330	0.090441	3.02	0.003	1.314334	(1.197886;1.573455)
3	0.318985	0.095220	3.35	0.001	1.375731	(1.138297;1.662691)
4	0.000000	0.000000	.	.	1.000000	(1.000000;1.000000)
source						
1	0.393438	0.092940	4.23	0.000	1.482067	(1.231856;1.783099)
2	0.523100	0.187570	2.79	0.007	1.687249	(1.161723;2.450507)
3	-0.197252	0.229756	-0.86	0.393	0.820984	(0.519764;1.296770)
4	-0.219548	0.136941	-1.60	0.113	1.245514	(0.948461;1.635601)
5	0.000000	0.000000	.	.	1.000000	(1.000000;1.000000)
sfed X age						
sfed=1	-0.562166	0.165546	-3.40	0.001	0.569973	(0.410024;0.792318)
sfed=0	0.000000	0.000000	.	.	1.000000	(10000000;1.000000)
sfed X age^2						
sfed=1	0.035024	0.009761	3.58	0.000	1.035645	(1.015726;1.055953)
sfed=0	0.000000	0.000000	.	.	1.000000	(1.000000;10000000)
sfed X age^3						
sfed=1	-0.000872	0.000239	-3.64	0.000	0.999128	(0.998653;0.999604)
sfed=0	0.000000	0.000000	.	.	1.000000	(1.000000;1.000000)
sfed X age^4						
sfed=1	0.000007	0.000002	3.53	0.000	1.000007	(1.000003;1.000011)
sfed=0	0.000000	0.000000	.	.	1.000000	(1.000000;1.000000)
sex X age						
sex=1	-0.019473	0.004011	-4.85	0.000	0.980715	(0.972920;0.988573)
sex=0	0.000000	0.000000	.	.	1.000000	(1.000000;1.000000)

R^2 = 0.0304 , $\hat{C}$=7.9427, df=8, p-value=0.4391

With so many interactions of breastfeeding with high order polynomial terms in age included in the model in Table 8.10, its interpretation becomes more difficult and it is necessary to plot the predicted probabilities to get a clearer understanding of the model.

Applying the theory of Section 8.7, one can estimate the probability of being malnourished from the model in Table 8.10. For example, consider the case of a female child with an uneducated mother in round 2 (June survey), whose source of income is from business income, over increasing age, separately for those children still breastfeeding and those not breastfeeding.

The two plots in Figure 8.2 based on the logits for the above conditions are:
For breastfeeders

$$g(y,x) = -4.038159 - 2.722320 + 1.348257(\text{age}) - 0.077343(\text{age}^2) - 0.001865(\text{age}^3) - 0.000015(\text{age}^4) - 0.828752 + 0.347920 + 0.27333 - 0.197252 - 0.562166(\text{age}) + 0.035024(\text{age}^2) - 0.000872(\text{age}^3) + 0.000007(\text{age}^4) - 0.019473(\text{age})$$

and for children not breastfeeding:

$$g(n,x) = -4.038159 + 1.348257(\text{age}) - 0.077343(\text{age}^2) - 0.001865(\text{age}^3) - 0.000015(\text{age}^4) - 0.828752 + 0.347920 + 0.27333 - 0.197252 - 0.019473(\text{age})$$

Therefore the probabilities of being malnourished for a specific age are:

$$P(\text{maln=yes}|\text{sfed=Yes,age=x}) = \exp(g(y,x)) / [1 + \exp(g(y,x))]$$

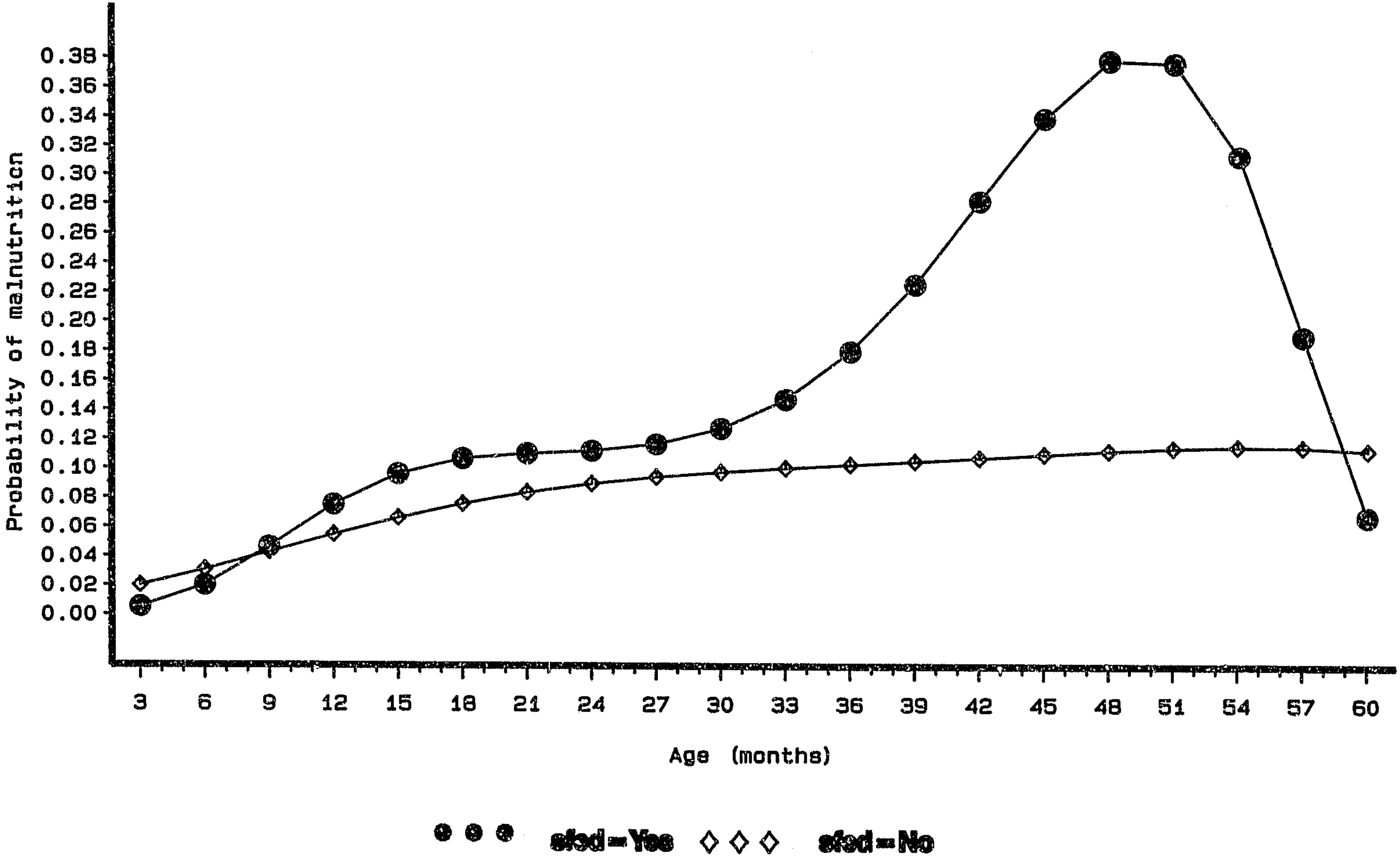
$$P(\text{maln=yes}|\text{sfed=No,age=x}) = \exp(g(n,x)) / [1 + \exp(g(n,x))]$$

The effect of the highly significant interaction terms (Table 8.10) between breastfeeding and age on the chosen levels of the independent variables, can be clearly observed by the plot in Figure 8.2. The 'still breastfeeding' group have lower probabilities of being malnourished than the 'not breastfeeding' group. For the first eight months the probability of malnutrition is higher among non-breastfeeders than breastfeeders. Thereafter there is a switchover, and for children that are breastfed, the probability of being malnourished increases sharply up until 48 months (0.378), and then decreases sharply to 0.068 at 60 months. For children that are not breastfeeding there is a gradual increase in the probability of malnutrition, increasing from 0.046 at 3 months to 0.112 at 60 months.

As there are numerous combinations of the predictor variables, each combination producing a different shaped graph, another approach is necessary to graphically represent the probability of malnutrition over increasing age in a more realistic way.

The output of SUDAAN produces predicted probability values for each observation. By adding the probability values in each age category and dividing by the number of individuals in that age category, the expected proportion in each age category is calculated. A plot of the expected proportion is seen in Figure 8.3 for all children under 5 years. For the first eight months the expected proportion of malnourished children is higher among non-breastfeeders than breastfeeders. This feature of the model is as expected since one would expect that breastfeeding is essential for healthy nutrition in the first year of life, and those children not receiving breastmilk have a higher probability of being malnourished. An unexpected result though is that after eight months breastfeeders have a higher expected proportion of malnourished children than non-breastfeeders. It would be more understandable and acceptable if the switchover occurred after 24 months when most children are no longer breastfeeding. This would confirm the hypothesis from Chapter 7 that those children still breastfeeding after 24 months come from poorer households

FIGURE 8.2 Plot of logistic regression model to predict malnutrition for all children under five years



where malnutrition is more common and where breastmilk is the main source of nutrition. The large variation seen in the expected proportions (see Figure 8.3) for children still breastfeeding above approximately 30 months is due to the small number of breastfeeders in these age categories and therefore uneven distribution of risk factors. A plot of the under 24 month model in Figure 8.4 (from Table 8.7) and the over 24 month model in Figure 8.5 (from Table 8.9) shows similar expected proportions for the same ages to those found in Figure 8.3. However, for the over 24 month model which does not have an age term, the expected proportion of being malnourished is fairly constant whereas in the overall model there is an increasing and then decreasing trend which is due to the quadratic, cubic and quartic terms in age being included in the model.

It was surprising to note from the model selection phases used to generate Tables 8.9 and 8.10 that place of residence was not a significant predictor of malnutrition. The reason for it not being included was due to its highly significant association with the mothers education level. To identify this association, education level was excluded as a predictor variable in the model, which resulted in place of residence being included and the coefficients of the other independent variables (i.e. source of income, sex, breastfeeding and age) hardly changing. With both variables considered in the model fitting process, place of residence was no longer significant. There was a better model fit with education included in the model than with place of residence included. These results suggest that the education level of the mother acted as an approximate 'surrogate' effect for place of residence. The highly significant association ($p<0.0001$) between place of residence and the education level of the mother can be seen in Table 8.11 where it is shown that rural mothers are significantly more uneducated than urban mothers.

FIGURE 8.3 Plot of the expected proportion malnourished in each age category for age < 5 years model

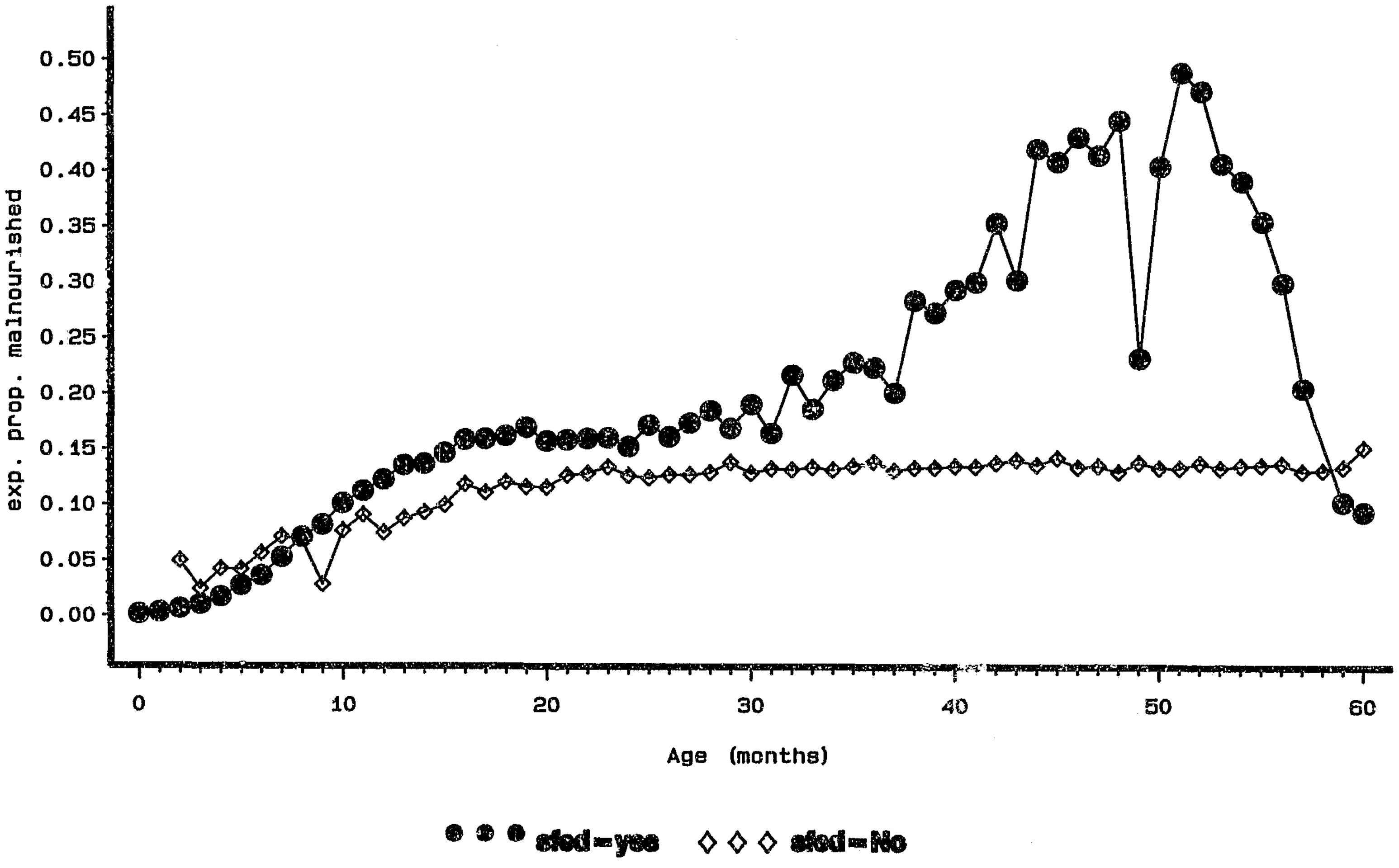


FIGURE 8.4 Plot of the expected proportion malnourished in each age category for age < 24 month

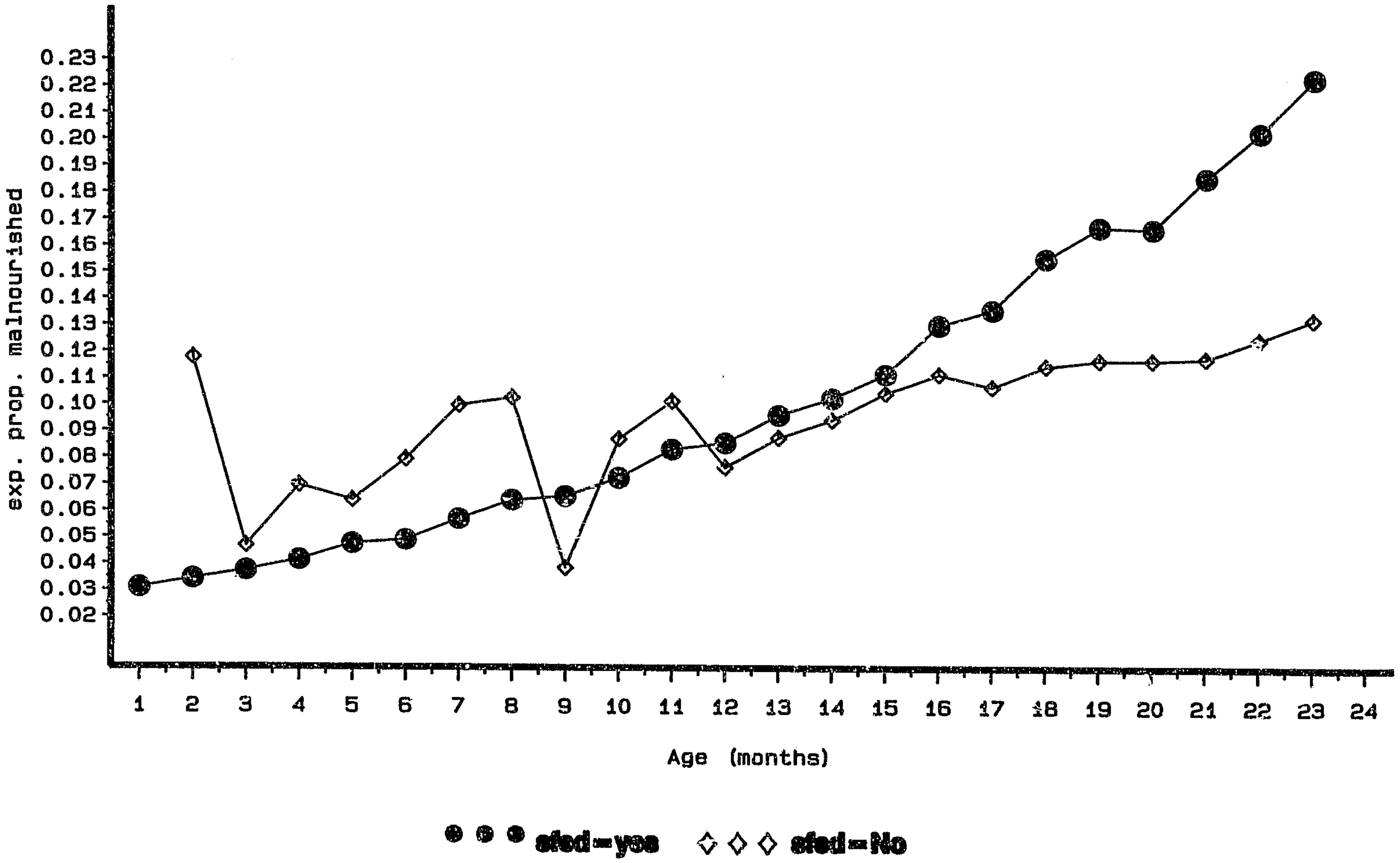


FIGURE 8.5 Plot of the expected proportion malnourished in each age category for age >= 24 month model

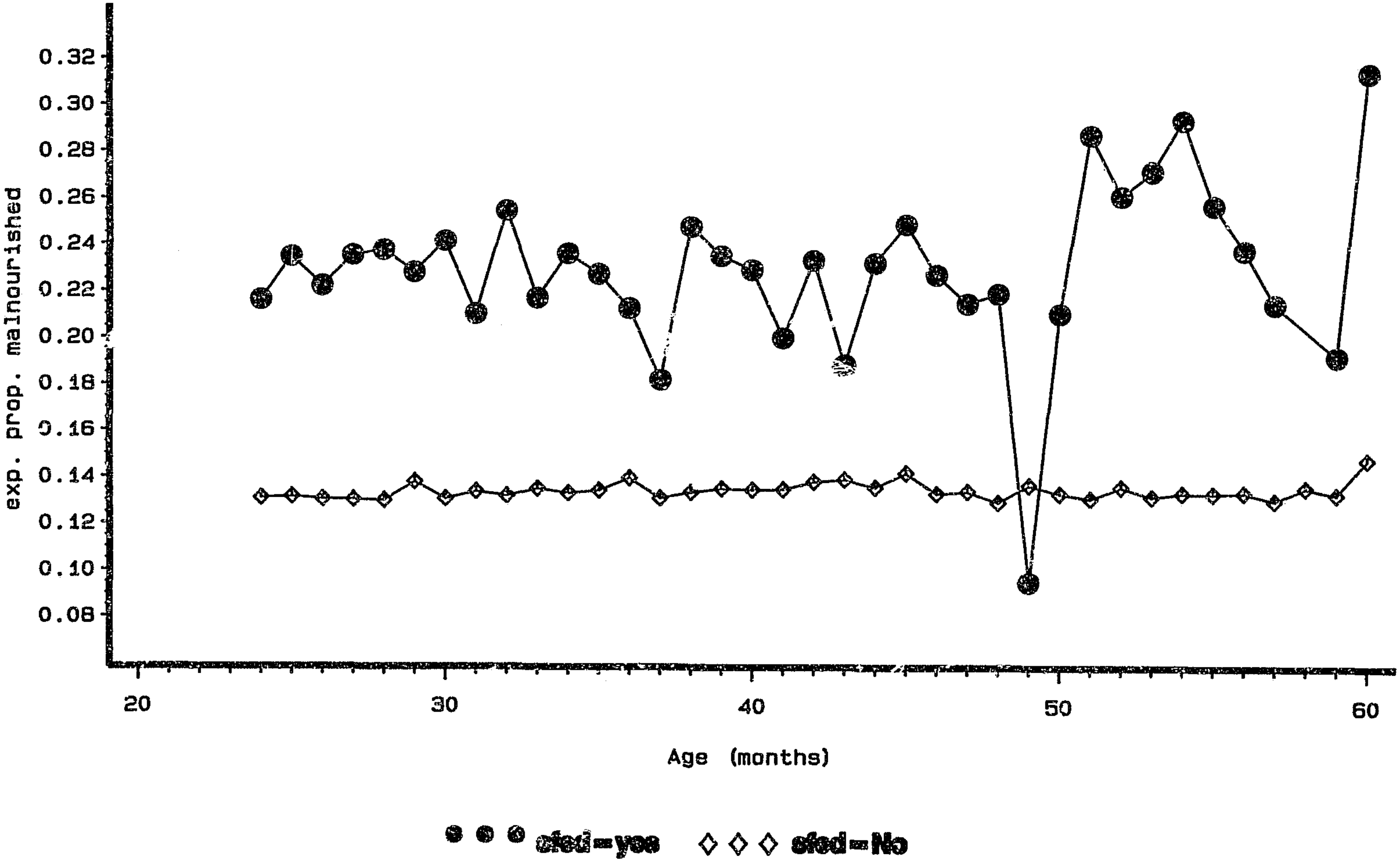


Table 8.11 Percentage of uneducated mothers by place of residence for all children under 5 years

	% (95 % CI)
Rural	76.0 (73.8 - 78.2)
Urban	30.8 (23.7 - 37.8)

$\chi^2_1 = 116.5$, p-value = 0.0000

To get a better understanding of the various factors influencing breastfeeding, a logistic regression model for breastfeeding was fitted in Table 8.12. An interesting result from Table 8.12 is that the variable rurb (place of residence) is a strong predictor of breastfeeding and that rural children are 4.8644 times more likely to breastfeed than urban children. A plot of the observed number of breastfeeders in rural and urban areas in Figure 9.1 of Chapter 9 confirms this result. The other predictor variables included in the model are age, the age X rurb interaction and quadratic and cubic terms in age. Even though there is a slightly positive age^3 effect, the stronger combined negative age and age^2 effects indicate a reduction in the odds of breastfeeding with increasing age. By substituting all possible age values in the model in Table 8.12, it was observed that the maximum probability for breastfeeding in rural (prob=0.993817) and urban (prob=0.970623) children were both at zero months (i.e. during the first month of life).

To further understand the association between breastfeeding, place of residence and malnutrition, a 2x2x2 log-linear model was also fitted to the under 24 month data in Table 9.8 of Chapter 9.

Table 8.12 Estimated Coefficients, Estimated Standard Errors, T-statistics, p-values, Odds Ratios and their confidenc intervals for the Multiple Logistic Regression Model to predict the probability of breastfeeding for children under 24 months (malnutrition excluded)

breastfed 1=Yes 0=No	Beta Coeff	SE Beta	T-test Beta=0	P-value T-test Beta=0	Odds Ratio	Odds Ratio 95% CI
constant	3.4977	0.369	9.470	0.000	33.041	(15.84;68.89)
age	-0.122	0.039	-3.120	0.002	0.885	(0.818;0.957)
age^2	-0.005	0.001	-3.496	0.001	0.995	(0.992;0.998)
age^3	0.000	0.000	5.929	0.000	1.000	(1.000;1.000)
rurb						
1	1.582	0.281	5.637	0.000	4.864	(2.783;8.501)
2	0.000	0.000	.	.	1.000	(1.000;1.000)
age X rurb	-0.041	0.011	-3.734	0.000	0.960	(0.939;0.981)

$\hat{C}=7.6688$, p-value=0.46647, $R^2=0.6710$

In explaining Figure 8.3, the trend in the first eight months goes against the strong breastfeeding association with malnutrition above eight months. In the first eight months children that are still breastfeeding have a lower chance of being malnourished than non-breastfeeders. The progressive increase in the expected proportion malnourished among the breastfeeders from 0.2 at 30 months to 0.5 at 51 months, can be explained in the same way as in Chapter 7. The reason is that children who are still breastfeeding at an older age tend to be from poorer households where breastmilk is still the main source of nutrition available, and where other nutritional supplements are not affordable. The breastmilk provided might not be sufficient or nutritious enough for the older child and therefore these children tend to have more malnutrition. The steep drop in this curve from 51 months to 60 months is probably due to the fact that there are very few children being breastfed at these ages whether they are from poorer households or not.

From the above analyses of the child nutrition data the following comments can be made:

(i) In the first eight months breastfeeders have a lower probability of being malnourished than non-breastfeeders, which goes against the general trend above eight months.

(ii) Children with mothers that were uneducated had higher levels of malnutrition. It was shown that the education level of the mothers acted as an approximate 'surrogate' effect for place of residence in predicting malnutrition. Therefore, for rural children, the beneficial effects of breastfeeding are outweighed by the negative effects of poverty and low levels of education of the mothers.

(iii) The children from households which draw primarily on migrant worker remittance are less likely to be malnourished than children of households where the primary source of income is from subsistence farming or the sale of livestock.

(iv) The log-linear model analysis is useful in confirming or clarifying associations between variables that are not clear in the logistic regression model.

(v) The single overall model is less parsimonious than the two separate models (below 24 months and above and equal to 24 months) to describe malnutrition in children, but it is useful to have a single model that does not use an arbitrary cut-off point to describe the process.

CHAPTER 9

RESULTS - MATERNAL CARE, DISABILITY, AND INJURY

9.1 INTRODUCTION

This chapter provides additional results from the NH&NS to those presented in Chapter 7 and Chapter 8 (dealing with child nutrition). These results are from the following four sections of the NH&NS: knowledge of health care facilities, maternal care, injuries, and disabilities. All the results in this Chapter are based on the techniques for analysing data from a complex survey design described in the earlier chapters of this thesis.

In a predominantly rural country such as Lesotho where hospitals and clinics are widely dispersed, a Village Health Worker(VHW) is extremely important in the treatment of injuries and illnesses and vital in the case of a medical emergency. The value of a VHW to the local community may be assessed at the simplest level, by the extent to which people are aware of the presence of a VHW in their area. The interviewer (supervisor) determined, for each community, whether a VHW existed or not. Table 9.2 relates the level of agreement between interviewers and respondents with regard to there being a village health worker (VHW) in the area of the interview or not using Cohen's Kappa.

A log-linear model is used to investigate how the education level of a mother affects both her breastfeeding habits and the likelihood of malnutrition in her child. A log-linear model is also fitted to assess the relationship between malnutrition, breastfeeding and place of residence (rural/urban) among children under 24 months. Of interest is whether the associations found in the logistic regression model in Chapter 8 persist for a log-linear model where the three variables are all explanatory.

The Maternal care section consists of various tables relating the use of ante-natal care and the place of delivery to factors such as place of residence, district, and past or present pregnancies. A Wald Chi-Squared test statistic (see Chapter 7) is calculated to determine whether associations exist between these variables. The probability of using ante-natal care and the probability of delivering at a clinic or hospital is predicted by fitting a logistic regression model to the data (see Chapter 8).

The structure of the question on disability required two sets of analyses. The reason being that three options, and therefore three variables, were given for a disability, with no order of importance provided (see Appendix E, Question 30). For example a respondent could have responded to the question 'Type of disability?' with : 'Speech problems', 'Amputation of 1 or more fingers' and 'Severe deafness' which would have resulted in disab1=m, disab2=a and disab3=l. The problem with this type of questioning is that it was unknown which disability was being referred to when the questions on the length of time of the disability (Question 31) and the cause of disability (Question 32) were asked. Two types of analyses for disability were therefore performed:

i) To get the prevalences of the different disabilities, 15 dichotomous variables were created, one for each disability type and if a disability appeared among the three variables in question 30 the value 1 was assigned to the corresponding variable, otherwise it would be left at zero. For example, the response disab1=m, disab2=a and disab3=l would assign a value of 1 to each of the three corresponding dichotomous variables. Using this approach, the disability rates by age, gender and place of residence were calculated for this chapter.

ii) The second approach used was based on the ranking of disability (according to an epidemiologist) from the most serious disability to the least serious i.e. out of the three possible disability responses for each respondent, only the most serious was considered. Using the same example as above the disability variable would be assigned the response '1' representing 'severe deafness' because the epidemiologist considered it to be the most serious of the three responses. With this approach a single cause was attributed to a disability and hence the two-way tables relating disability to cause could be presented in Tables 9.30 and 9.31. It should be noted, however, that only 6.6% of those with any disability reported more than one disability.

A logistic regression model is fitted to the data to predict the probability of disability from occupation, source of income, gender, age, place of residence, and education level.

Finally tables are presented of injuries cross-tabulated by gender, place of residence, and age group.

The results presented in this Chapter and in the previous few chapters on child nutrition should provide an epidemiologist with valuable information on child nutrition, breastfeeding patterns, maternal care, disabilities and injuries. Interventions such as feeding and education programs and the provision of health care facilities can be evaluated, based on, for example, the levels of malnutrition among children and the prevalence of the different types of disabilities and injuries in the different areas of Lesotho.

9.2 KNOWLEDGE OF HEALTH CARE FACILITIES

In order to evaluate the level of agreement between interviewers and respondents with regard to there being a village health worker (VHW) in the area of the interview or not, Cohen's Kappa is calculated. Consider the 2x2 table in Table 9.1 indicating the responses of two raters to a series of yes/no questions. The proportions where the raters are in agreement are a and d, and the proportions where they are in disagreement are b and c. The total proportion agreed by both raters is therefore p=a+d. Kappa has similar properties to the correlation coefficient, namely when there is complete agreement Kappa takes on the value 1, when there is complete disagreement then Kappa takes on -1. For multinomial sampling, the sample statistic $\hat{K}$ has a large-sample normal distribution (see Agresti (1990) p.366).

The formula for Kappa is

$$\hat{K} = \frac{2(ad + bc)}{p_1q_2 + p_2q_1}$$

<u>Table 9.1</u> **A 2x2 contingency table of the proportions of agreement and disagreement used in the definition of Cohen's Kappa statistic**

	Rater B		
Rater A	Response=Yes	Response=No	Total
Response=Yes	a	b	p_1
Response=No	c	d	q_1
Total	p_2	q_2	1

Table 9.2 Percentage of agreement between interviewers and respondents based on the presence or absence of Village Health Care Workers in the area of the respondent

	Rural (Total percent)		Urban (Total Percent)	
	Respondent VHW=Yes	Respondent VHW=No	Respondent VHW=Yes	Respondent VHW=No
Interviewer VHW=Yes	52.0 (44.2-59.7)	0.2 (0.1-0.3)	6.2 (0.0-13.7)	0.3 (0.0-0.6)
Interviewer VHW=No	0.9 (0.4-1.3)	46.9 (39.1-54.7)	0.6 (0.0-1.5)	92.9 (85.6-100.0)
	Kappa=0.978 Z=4.374 p=0.000		Kappa = 0.928 Z=4.149 p=0.000	

Table 9.2 presents the responses of two raters to the yes/no question regarding the presence or absence of a VHW in the area by the rural/urban split. From this table it is clear that there is a high level of agreement between the interviewer and respondent both in the rural and urban areas of Lesotho. If it is assumed that the interviewer's knowledge as to the presence or absence of a village health worker (VHW) is correct, then both rural and urban respondents are also highly aware of the VHW's presence in their areas. This result should be encouraging for the Lesotho Ministry of Health (LMOH) in that the population appears to be well-informed of the existence of Village Health Care Workers in the different areas. Of concern to the LMOH from Table 9.2 should be the high percentage (47.8%) in the rural areas without a VHW.

9.3 BREASTFEEDING

The associations between breastfeeding, malnutrition, and the education level of mothers of children in their weaning years (under 2 years) are of interest to the LMOH. In order to determine the associations simultaneously, a log-linear model is applied to the data. The SAS program STANDERR that has been written for this thesis (from 7.6.1 of Chapter 7) using Fellegi's correction factor is used for this analysis.

The data on which this model has been fitted is given in Table

9.3, and the results are given in Table 9.7.

Table 9.3 Three-way table of education levels of mothers by breastfeeding and by nutritional status of children under two years

sample size adjusted sample weighted sample design effect	Malnourished=No		Malnourished=Yes	
	Uneducated	Educated	Uneducated	Educated
Still Breastfed	1907 1993 44550 1.51	856 778 17391 1.45	225 226 5047 1.37	64 62 1385 1.45
Not Breastfed	379 402 8975 1.24	212 183 4098 1.45	58 62 1386 1.19	26 20 458 1.02

The saturated model for the above table is the same as that given in equation 7.4.2 of Chapter 7, where A refers to the education level of the mother, B indicates whether the child is still breastfed or not, and C indicates whether the child is malnourished or not.

The backward model selection technique described in Section 7.4.2, has been used to determine the best log-linear model for the above data. It should be noted that G^2 given in Tables 9.4 and 9.5 has been divided by the average design effect.

Table 9.4 Probability that the saturated model less the three-way interaction term is adequate

Effect	DF	G^2	P-value
base model	0	0.0000	1.000
ABC	1	0.0007	0.979

The base model given in the first row of Table 9.4 is the saturated model which fits the data perfectly, and therefore there are no degrees of freedom to fit the model. The second row is the model without the three-way interaction term ABC. The incremental effect of ABC is examined by comparing the differences between the G^2 for the saturated model and the model without the ABC term. Thus the value for G^2 for the incremental effect is 0.0007 - 0.0000 = 0.0007, which is not significant (p-value=0.979). Therefore the three-way interaction does not make a significant contribution to the model and can be eliminated.

To further simplify the model, the possibility of deleting one of the pairwise interactions is tested in Table 9.5 and Table 9.6. The base model is now the model containing all two-way interaction terms and main effects, but no three-way interaction term.

Table 9.5 Testing the model adequacy by excluding two-way interaction terms

Effect	DF	G^2	P-value
base model (excluding ABC)	1	0.0007	0.9790
BC	2	2.0723	0.3548
AC	2	5.8075	0.0549
AB	2	3.9596	0.1381

Table 9.6 The incremental effects of excluding the two-way interaction terms

Effect	DF	G^2	P-value
BC	1	2.0716	0.1506
AC	1	5.8068	0.0160
AB	1	3.9589	0.0466

The results in Tables 9.5 and 9.6 indicate that the BC interaction can be eliminated from the model. However, the other two interaction terms AC and AB need to be included. These results indicate that there is a significant association between the education level of the mother and nutritional status of the child, and between nutritional status of the child and regression model to the data of children under 24 months, the place of residence (rural/urban) was a significant predictor of breastfeeding i.e. children from a rural area were more likely to be breastfeeding. To get a better understanding of the underlying relationships between breastfeeding, malnutrition and place of residence, a log-linear model shown in Table 9.9 was also fitted to the under 24 month data given in Table 9.8.

Table 9.8 Three-way table of place of residence by breastfeeding and by nutritional status of children under 24 months

sample size adjusted sample weighted sample design effect	Still breastfed=Yes		Still breastfed=No	
	Normal	Malnourished	Normal	Malnourished
Rural	2349 2529 56512 1.61	262 271 6057 1.61	391 436 9740 1.50	56 60 1337 1.36
Urban	430 256 5732 1.81	25 15 340 1.65	125 76 1704 1.88	12 7 146 0.84

The main effects in Table 9.9 indicate that more children are being breastfed than not, there are more rural than urban children in the sample and there are less malnourished than adequately nourished children. A higher than average number of children under 24 months that are malnourished and still breastfeeding are from rural areas, whereas children who are adequately nourished and not breastfeeding are from urban areas.

It is interesting to note that the association between malnutrition and breastfeeding is not significant in this model. The reason for this result can be explained by the logistic regression model of Chapter 8 (see Table 8.10) where a significant association between breastfeeding and age was observed for children under 24 months. The above log-linear model however does not take into account this age effect. The absence of the place of residence by nutritional status association can also be explained by Table 8.10, where the place of residence was not a significant predictor of malnutrition (for children under 24 months).

Table 9.9 Estimated model coefficients resulting from the selected model

Factor	Coefficient	Std. Err.	Z-Value	95% CI
Constant	4.72958	0.06245	75.7288	(4.607;4.852)
$\hat{\lambda}_1^A$	0.72852	0.06245	11.6649	(0.606;0.851)
$\hat{\lambda}_2^A$	-0.72852	0.06245	-11.6649	(-0.851;-0.606)
$\hat{\lambda}_1^B$	1.03075	0.06245	16.5041	(0.908;1.153)
$\hat{\lambda}_2^B$	-1.03075	0.06245	-16.5041	(-1.153;-0.908)
$\hat{\lambda}_1^C$	1.11715	0.06245	17.8875	(0.995;1.240)
$\hat{\lambda}_2^C$	-1.11715	0.06245	-17.8875	(-1.240;-0.995)
$\hat{\lambda}_{11}^{AB}$	0.13688	0.06245	2.19169	(0.014;[illegible])
$\hat{\lambda}_{12}^{AB}$	-0.13688	0.06245	-2.19169	(-0.259;0.014)
$\hat{\lambda}_{21}^{AB}$	-0.13688	0.06245	-2.19169	(-0.259;0.014)
$\hat{\lambda}_{22}^{AB}$	0.13688	0.06245	2.19169	(0.014;0.259)

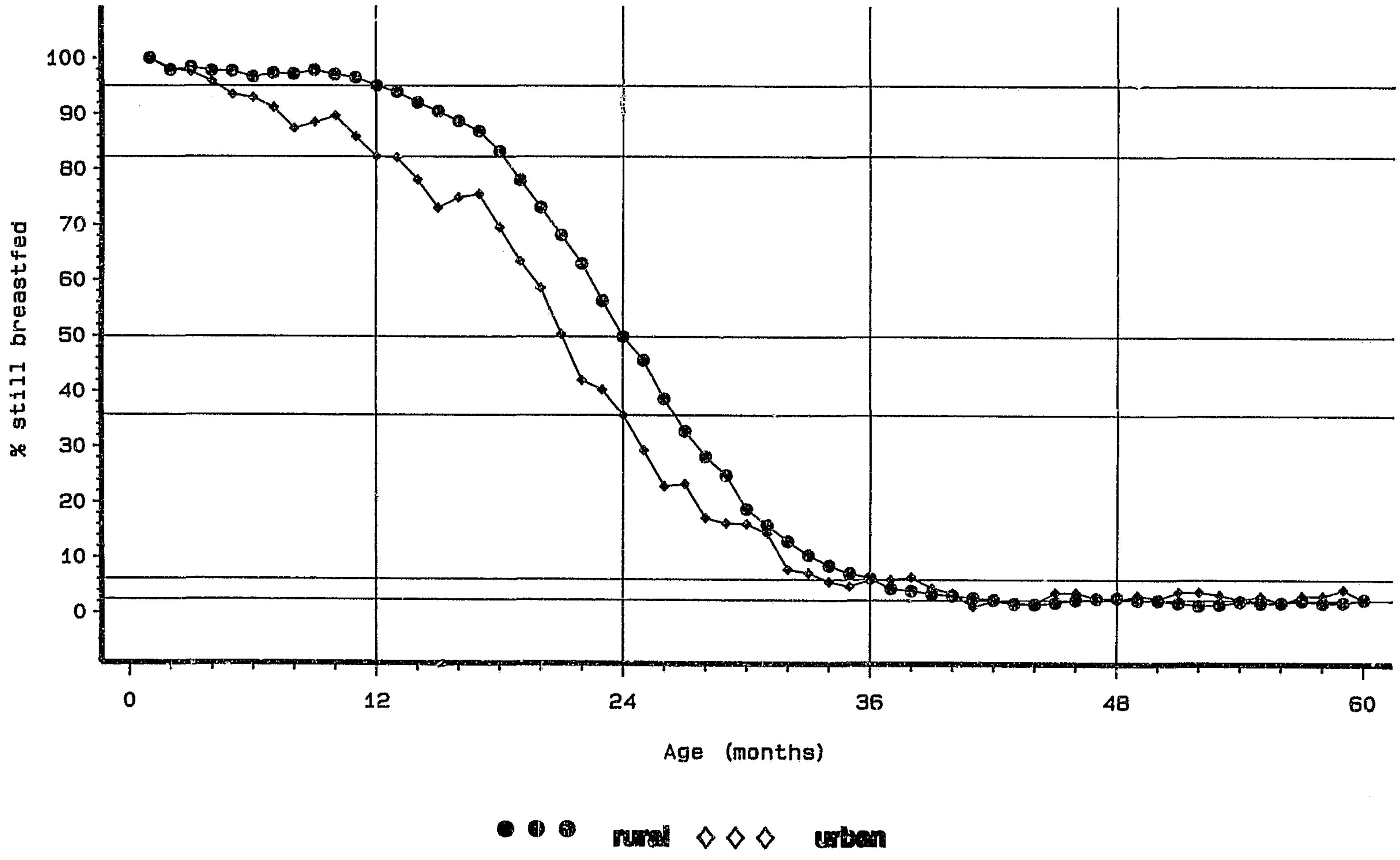
A = 1 ... still breastfed
A = 2 ... not being breastfed

B = 1 ... rural
B = 2 ... urban

C = 1 ... not malnourished
C = 2 ... malnourished

A plot of the breastfeeding habits of the Lesotho population by place of residence can be seen in Figure 9.1. The five point moving average plot indicates a similar trend for both the rural and urban areas although between 8 and 26 months there are consistently 15% more children still breastfeeding in the rural areas than in the urban areas. After 36 months the proportions remain almost the same.

FIGURE 9.1 Five month moving average plot of the percentage of children still breastfeeding by age and place of residence



9.4 MATERNAL CARE

Ante-natal care is directed at optimising the outcome of the pregnancy for both mother and foetus during the course of the pregnancy and the perinatal period. This can be expected to have far-reaching implications for the prevention of disabilities which may affect the newborn for the rest of its life.

This is ideally achieved by an ongoing programme of monitoring for early detection of current and future problems, as well as the active management of existing pathology. In addition, maternal training and education can be implemented during ante-natal visits, for better pre and post-natal care.

For example, women with inadequate pelvic dimensions for delivery, or foetal malpositions can be identified early, and appropriate delivery techniques such as caesarian section can be planned, or corrective measures instituted.

In addition to this, aspects of maternal health which may be exacerbated by pregnancy, can be identified and managed.

9.4.1 ANTE-NATAL CARE

This section considers ante-natal care, i.e. the distribution of the population using ante-natal care (ANC), the places used for ANC, the number of visits for ANC, the reasons for not using ANC, and finally a prediction of ANC using a logistic regression model.

From Table 9.10 and 9.11, it can be seen that the majority of women attended ANC in their last pregnancy (89.6%) and that there is no significant difference between the rural and urban previously pregnant women in their attendance of ANC (yes or no) $(p>0.05)$.

Table 9.10 **Percentage of women who attended ante-natal care (ANC) in their last pregnancy (excludes currently pregnant)**

ANC	% (95% CI)
Attended	89.6 (87.8-91.3)
Did not attend	10.4 (8.7-12.2)

N = 95471 n = 4892 (N represents the number weighted to the population and n represents the sample size)

Table 9.11 **Percentage of women who attended ante-natal care (ANC) by place of residence (excludes currently pregnant)**

	% (95% CI)	N
Rural	89.1 (87.2-91.1)	83634
Urban	92.4 (89.2-95.6)	6324

χ^2_1 = 3.13 , p-value = 0.081

The results in Table 9.12 indicate that when pregnant women are also considered in the analysis there is a significant difference between the rural and urban areas' usage of ANC. Women with livebirths and pregnant women who have attended ANC mostly go to a clinic (87.37%). Some of the reasons for non-attendance are given in Tables 9.17 and 9.18.

Table 9.12 Distribution of women, who in their last pregnancy or current pregnancy, attended ante-natal care (ANC) by source of care

	Rural % (95% C.I.)	Urban % (95% C.I.)	Total % (95% CI)
Clinic	86.86 (85.1-88.6)	91.04 (88.1-94.0)	87.37 (85.8-89.1)
Private doctor	0.68 (0.3-1.0)	2.01 (0.6-3.4)	0.84 (0.5-1.1)
Traditional	0.27 (0.2-0.4)	0.21 (0.0-0.6)	0.27 (0.1-0.4)
Trained VHW	0.31 (0.2-0.4)	0.18 (0.0-0.4)	0.29 (0.2-0.4)
Self Care	0.61 (0.3-1.0)	0.38 (0.0-0.7)	0.59 (0.3-0.9)
Other	0.21 (0.1-0.3)	0.0	0.18 (0.0-0.3)
Did not attend	11.06 (9.4-12.7)	6.18 (4.0-8.4)	10.46 (9.0-11.9)
N	96151	13361	109512

χ^2_6=28.49, p-value=0.0003

To measure the use of "Western style" care (i.e. use of a clinic or doctor), a two-way table of pregnant women who received ANC from the clinic or doctor by place of residence is shown in Table 9.14. There is a significantly smaller proportion of rural women than urban women who were pregnant and received ANC from the clinic or doctor at the time of the survey (p=0.001).

Table 9.13 Distribution of pregnant women at the time of the survey who attended ANC

	% (95% CI)
Attended ANC	65.0 (61.5-68.4)
Did not attend ANC	35.0 (31.5-38.4)

N = 19309

Currently pregnant women have a higher percentage of not attending ANC (35%) (Table 9.13) than previously pregnant women (10.4%) (Table 9.10). This seemingly surprising result can be explained by the truncation effect. From Figure 9.4 it is clear that the month (duration) of the pregnancy has a strong effect on whether the pregnant women attends ante-natal care or not. Therefore in a cross-sectional survey of currently pregnant women, a sizeable proportion will still be at an early phase of their pregnancies and will not yet have started attending ANC. For example a pregnant women in her third month of pregnancy might have responded 'Did not attend ANC' at the time of the survey, but was planning to attend ANC in her fifth month of pregnancy.

Table 9.14 Distribution of pregnant women at the time of the survey who received ANC from the clinic or doctor by place of residence

	Clinic or Doctor	
	% (95% CI)	N
Rural	62.4 (58.8-66.1)	10708
Urban	77.3 (71.2-83.4)	2071
Total	64.0 (60.7-67.4)	12779

$\chi^2_1 = 16.0$, p-value = 0.0001

In Table 9.15 a significant association is seen to exist between the 10 districts of Lesotho and the percentage of pregnant women who received ANC from the clinic or doctor i.e. the use of clinics or doctors by pregnant women varies significantly amongst the different districts. Of interest is the high percentage (73.5%) of women in the predominantly rural district of Thaba-Tseka that received ANC from a clinic or doctor and the relatively low percentage in the mixed (rural and urban) districts of Berea (58.2%) and Leribe (58.3%). which goes against the overall trend noted in Table 9.14. These differences are partly due to the uneven distribution of doctors and clinics across Lesotho.

Table 9.15 Percentage of pregnant women at the time of the survey who received ANC from the clinic or doctor by district

District	% (95% CI)	N
Butha-Buthe	72.4 (57.4-87.3)	1286
Leribe	58.3 (51.1-65.5)	3294
Berea	58.2 (46.8-69.5)	2192
Maseru	70.3 (63.3-77.3)	3993
Mafeteng	70.8 (58.9-82.8)	2070
Mohale Hoek	62.5 (49.5-75.4)	2107
Quthing	47.9 (37.0-58.8)	1786
Quacha Nek	63.5 (42.1-84.9)	658
Mokhotlong	60.5 (55.2-65.8)	759
Thaba-Tseka	73.5 (60.6-86.4)	1164
Total		19309

$\chi^2_9 = 27.79$, p-value = 0.0031

Table 9.16 shows that in their last completed pregnancy most women (59.9%) made between three and six visits for ANC, with a small percentage only making one visit, and a relatively high percentage making nine or more visits. This trend is further illustrated by the cumulative percentage of number of visits to ANC (those mothers that did not attend ANC are also included) in Figure 9.2. It is shown that 72.65% of women attended ANC less than six times, which, according to Schneider, Mavrandonis and Price (1991) is the suggested minimum target for attendance of ANC. Twelve percent of women made only one visit to ANC.

Figure 9.2 Percentage of women who made x or fewer visits to ANC during last completed pregnancy

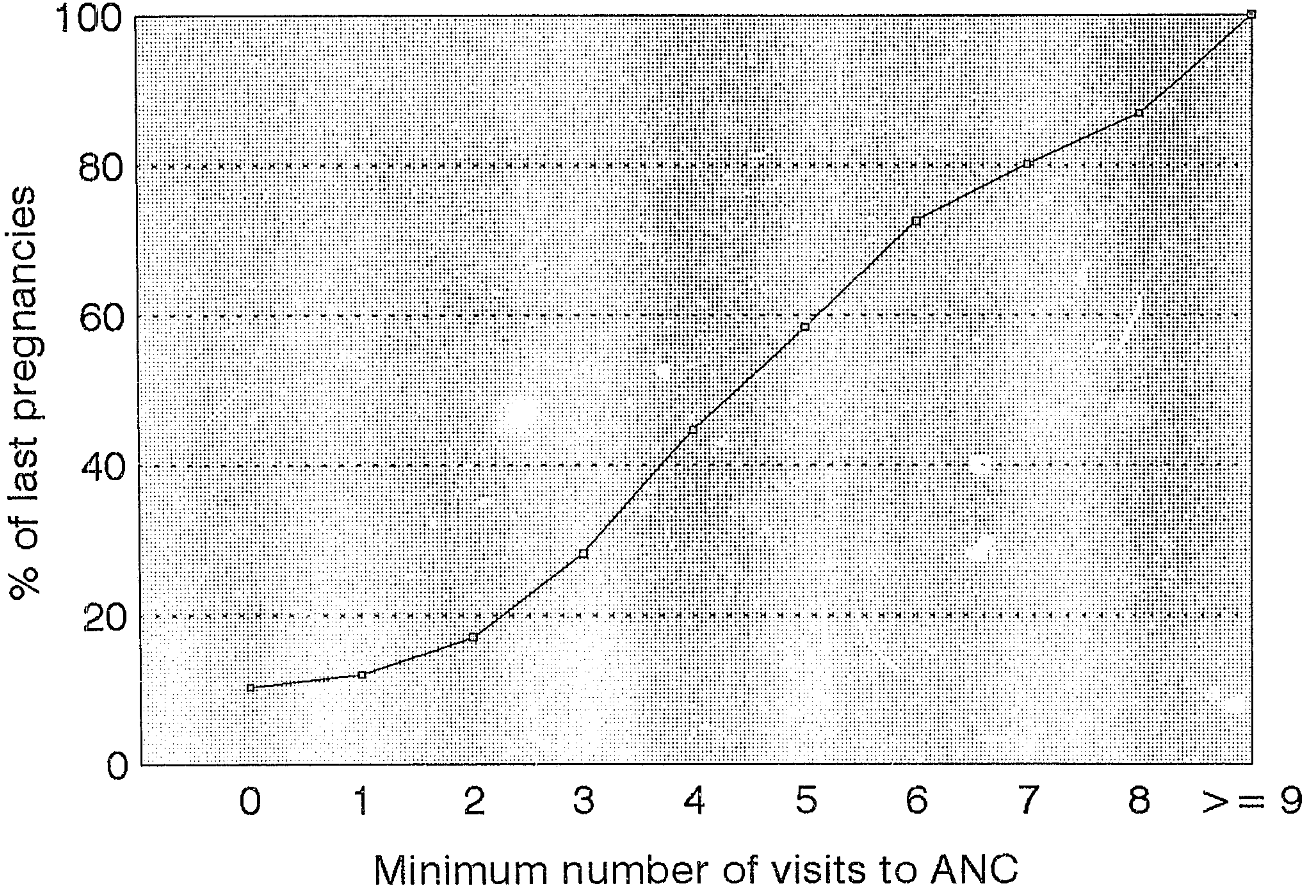


Table 9.16 Distribution of the reported number of visits for ANC made by women during their last pregnancy (excluding currently pregnant woman)

No. of visits	% (95% CI)	N
1	1.9 (1.3-2.4)	1585
2	5.5 (4.4-6.5)	4587
3	13.4 (11.9-14.9)	11176
4	17.2 (15.2-19.1)	14345
5	15.1 (13.4-16.8)	12594
6	15.7 (13.8-17.6)	13094
7	8.3 (6.9-9.6)	6922
8	7.6 (6.3-8.9)	6338
9 or more	15.4 (11.8-19.1)	12843
Total	100.0	83402

(Note: 10% did not attend ANC at all.)

As is expected, Figure 9.3 shows that the longer the duration of the pregnancy the larger are the average number of visits for ANC. Figure 9.4 shows that there is a steady increase in the percentage of women who made at least one visit to ANC from the first month of pregnancy (23%) to almost all women in the last month of pregnancy (93%). The relatively lower percentage of attendance in the first three months of pregnancy can either be attributed to women not being aware of their pregnancy or not feeling it necessary to attend ante-natal care at such early stages of pregnancy.

Figure 9.3 The average number of visits for Ante-Natal Care by duration of pregnancy

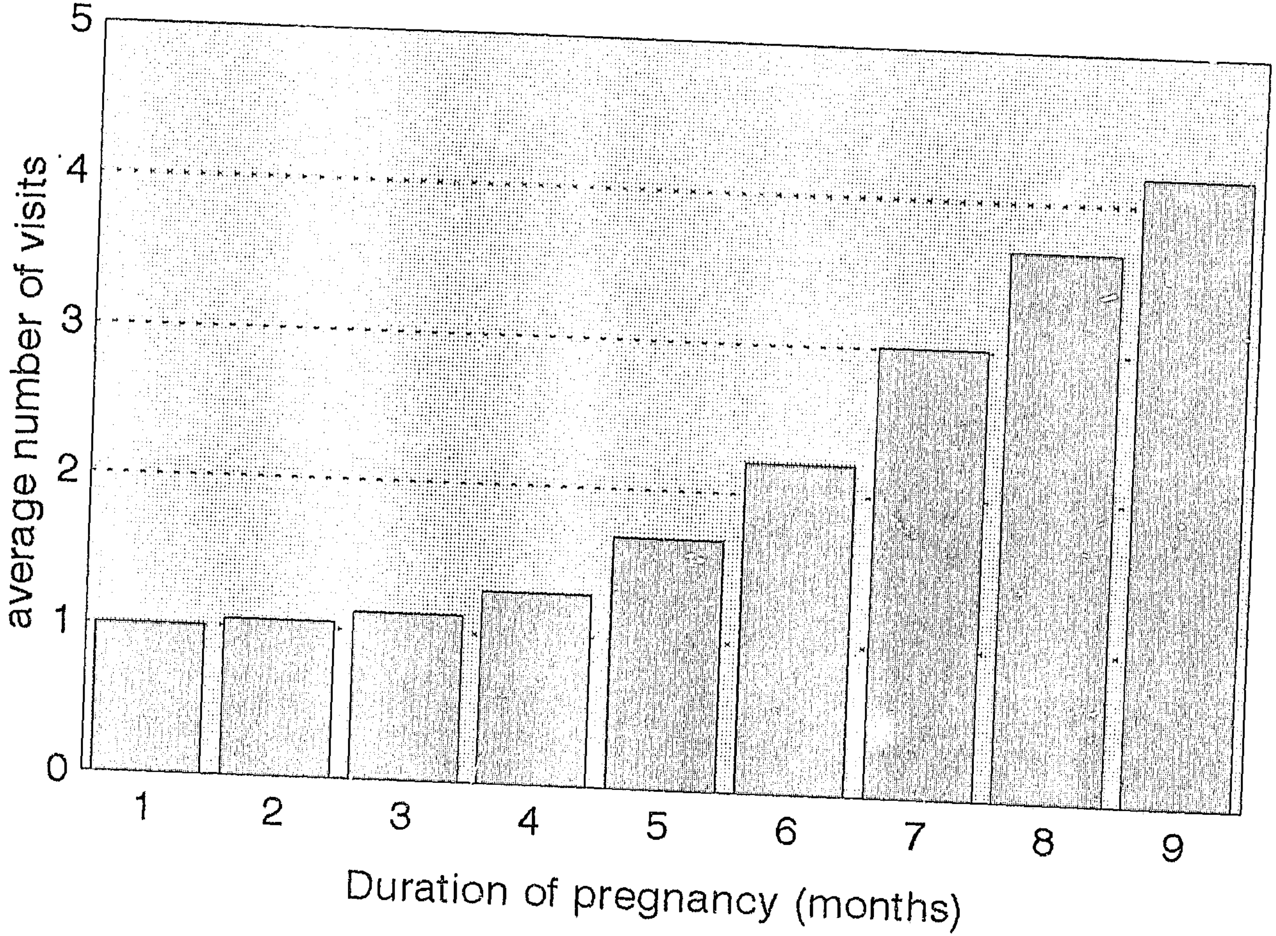


Figure 9.4 Percentage of women who have made at least one visit to ANC by duration of pregnancy

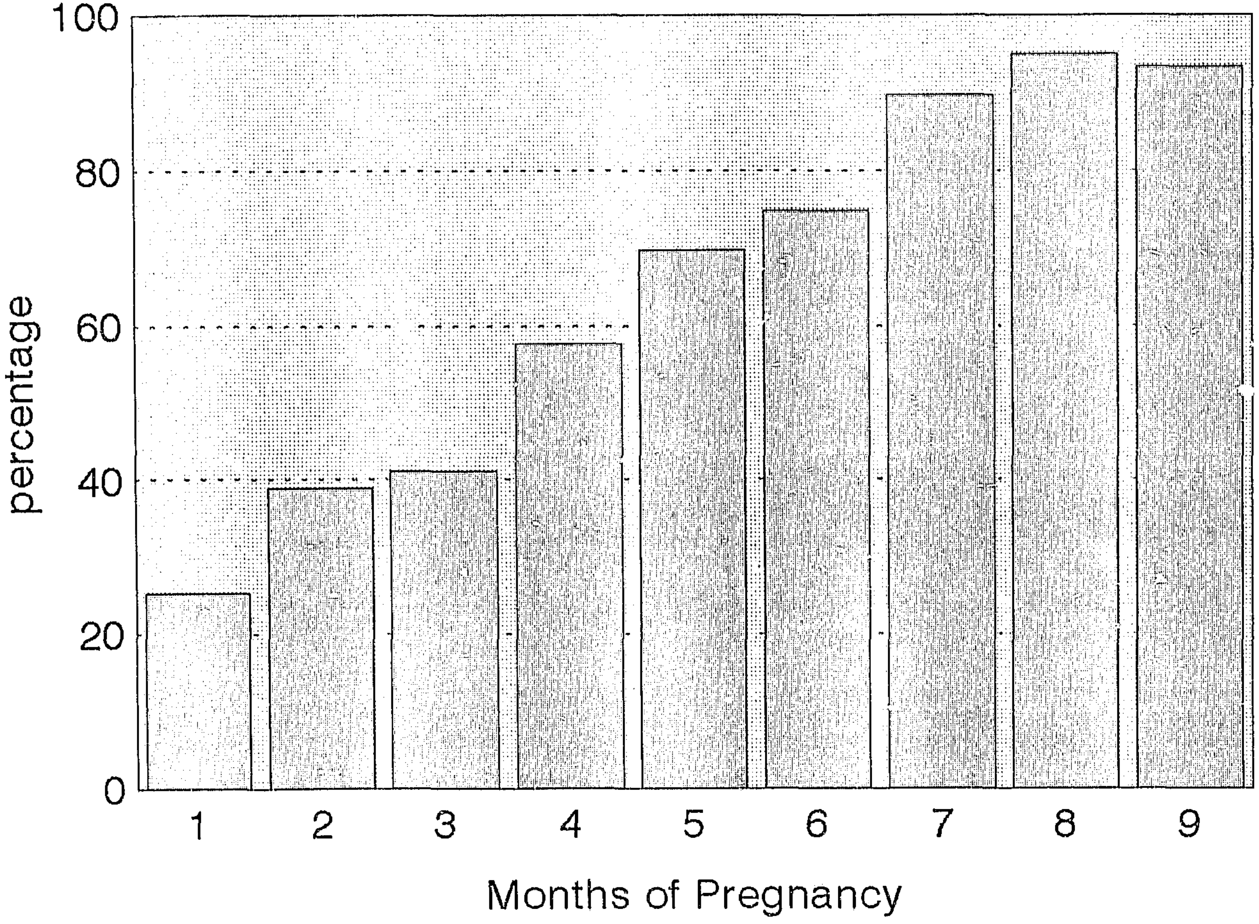


Table 9.17 Distribution of the reasons given by women for non-use of a clinic or doctor for ANC during past or present pregnancy by place of residence

	Clinic too far % (95% CI)	No money % (95% CI)	Witchdoctor % (95% CI)	Other % (95% CI)	No reason % (95% CI)	N
Total	9.1 (6.0-12.2)	33.3 (29.2-37.4)	3.6 (1.95-5.3)	44.3 (39.0-49.6)	9.7 (6.7-12.6)	11194
Rural	9.0 (5.7-12.3)	34.2 (29.9-38.5)	3.4 (1.65-5.1)	43.7 (38.3-49.1)	9.7 (6.6-12.8)	10402
Urban	10.3 (0.8-19.8)	20.9 (7.2-34.6)	7.0 (0.7-13.4)	51.4 (28.4-74.4)	10.2 (2.4-18.0)	792

$\chi^2_4 = 7.95$, p-value = 0.106

No significant association exists between place of residence and the reasons for non-use of a clinic or doctor in Table 9.17. This surprising result is due to the effect of clustering and stratification because if the data from this two-way table were considered to come from a simple random sample then χ^2_4=79.65 which is highly significant ($p<0.0001$). Of interest to health planners is the high proportion of respondents who 'have no money' as their reason for not using a clinic or doctor. The questionnaire however, did not distinguish the lack of money for consultation from the lack of money for transport.

From Table 9.18 there is a highly significant difference between districts in the reasons for non-use of a clinic or doctor. The reason 'Clinic too far', was given by the highest percentage (27.4%) of women who did not go to a clinic or doctor for ANC in the rural district of Thaba-Tseka. It should be noted though that there were a small number of responses (N=472, n=20) in this district and hence the wide confidence interval. The reason 'No Money' was given by high percentages of women in the districts of Mafeteng(45.8%), Quachas Nek (36.7%) Thaba Tseka(36.5%) and Berea(41.4%), whereas in Mokhotlong only 14.8% of those who did not attend ANC cited this as their reason.

Table 9.18 Distribution of the reasons of all pregnancies (current and past) for non-use of a clinic or doctor for ANC by district

	Clinic too far % (95% CI)	No money % (95% CI)	Witchdoctor % (95% CI)	Other % (95% CI)	No reason % (95% CI)	n	N
Butha-Buthe	0.0	29.1 (14.4-43.8)	0.0	43.5 (6.9-80.1)	27.4 (5.4-[illegible].4)	18	431
Leribe	5.65 (2.1-9.[illegible]	27.8 (18.2-37.4)	3.79 (0.1-7.5)	53.3 (40.9-65.6)	9.4 (4.1-14.65)	89	1810
Berea	4.1 (0.1-8.1)	41.4 (30.3-52.5)	3.13 (0.0-7.4)	47.2 (36.7-57.8)	4.8 (0.0-9.2)	64	1363
Maseru	7.9 (1.9-13.9)	30.9 (22.9-38.9)	3.27 (0.0-6.8)	48.1 (35.9-60.1)	15.6 (9.8-21.4)	87	1712
Mafeteng	5.7 (1.1-10.3)	45.8 (36.7-54.9)	2.27 (0.0-5.9)	37.1 (26.6-47.6)	16.5 (9.2-23.8)	90	1961
Mohale Hoek	11.9 (0.0-23.8)	31.4 (14.5-48.2)	3.41 (0.0-7.8)	46.6 (31.6-61.5)	6.7 (0.0-14.7)	61	1135
Quthing	16.6 (6.8-26.4)	27.6 (18.5-36.7)	3.69 (0.0-8.4)	37.5 (16.7-58.2)	14.6 (2.0-27.2)	55	1253
Quacha Nek	0.0	36.7 (34.0-39.4)	9.3 (0.0-26.5)	51.9 (32.5-71.3)	2.0 (0.0-5.96)	22	434
Mokhotlong	22.0 (7.1-36.8)	14.8 (6.7-22.9)	10.9 (2.0-19.7)	37.1 (15.2-59.0)	15.1 (0.8-29.4)	28	622
Thaba-Tseka	27.4 (0.0-59.7)	36.5 (10.2-62.8)	0.0	29.7 (0.00-63.0)	6.4 (0.3-12.5)	20	472
Total						534	11194

χ^2_{36} = 5364 , p-value = 0.000

To predict ante-natal care (ANC=1 indicates usage, ANC=0 indicates non-usage) a multiple logistic regression model was fitted in Table 9.19. The predictor variables fitted in this model are place of residence, age, occupation level, number of previous births, education level and source of income.

Table 9.19 Multiple Logistic Regression Model to predict Ante-Natal Care (ANC) (current and last pregnancy)

ANC	Degrees of freedom	S_waite Adj DF	S_waite Adj Chisq	p-value S_waite Adj Chisq
overall model	7	5.01	986.86	0.000
model minus intercept	6	4.97	44.18	0.000
rural/urban	1	1.00	3.23	0.073
source of inc[illegible]	4	3.58	17.93	0.000
education le[illegible]	1	1.00	16.53	0.000

X^2=11.8986 df=7, p-value=0.10394

R^2=0.01249

The Pearson chi-square (p-value=0.10394) indicates there is not a significant lack-of-fit of the model in Table 9.12. Only two effects (education level and source of income) proved to be significant in predicting the attendance of ante-natal care. Place of residence, age, occupation level, and number of past pregnancies were not significant and were therefore not included in the model.

<u>Table 9.20</u> Estimated Coefficients, Standard Errors, T-tests, and Odds Ratios for the Multiple Logistic Regression Model determining the use of Ante-Natal Care

ANC 1 = attended 0 = did not attend	Beta Coeff	SE Beta	T-test Beta=0	P-value T-test Beta=0	Odds Ratio	Odds Ratio 95% CI
Constant	2.48	0.15	16.42	0.000	11.96	(8.66;16.16)
Uneducated	-0.50	0.12	-4.30	0.000	0.61	(0.48;0.76)
Educated	0.00	0.00	.	.	1.00	(1.00;1.00)
Source of Income						
1	-0.44	0.16	-2.86	0.0054	0.64	(0.47;0.87)
2	-0.38	0.22	-1.76	0.0820	0.68	(0.44;1.05)
3	0.26	0.34	0.77	0.4456	1.30	(0.66;2.56)
4	0.08	0.18	0.43	0.6716	1.08	(0.75;1.55)
5	0.00	0.00	.	.	1.00	(1.00;1.00)

Education level: Educated = 'Standard 5 or above education level', Uneducated = 'below Standard 5 education level',
Source of income: 1='Subsistence farming', 2='Cash-cropping/sales of livestock', 3='Business Income', 4='Cash remittance from migrant workers', 5='Wages/salaries in cash'.

The above logistic regression model predicts that an uneducated pregnant woman is 0.61 times less likely than an educated pregnant woman of attending ANC. Relative to Wages/salaries in cash, those females with a source of income from subsistence farming are less likely to attend ANC (an odds ratio of 0.64), while cash-cropping/sales of livestock, business income and cash remittance from migrant workers do not affect the probability of attending ANC.

9.4.2 DELIVERY

Factors affecting birth delivery are considered in this section. These include the place of delivery, the reasons for not using a clinic or hospital and who assisted the delivery. For current pregnancies the "intended" place of delivery is used. The place of delivery of past pregnant women is predicted using a logistic regression model.

9.4.2.1 CURRENT PREGNANCIES

Table 9.21 shows there to be a significant association between the place of residence and the place of delivery intended by currently pregnant women. In the rural areas a smaller percentage of the women (73.8%) intend to use the hospital than in the urban areas (92.8%), and 14.2% of the rural women intend using a clinic compared to only 2.4% of urban women. A relatively high proportion of women (10.7%) in the rural areas intend using the home as a place of delivery, as opposed to only 3.9% of the urban women.

Table 9.21 Distribution of the place of delivery intended by currently pregnant women by place of residence

Place of delivery	Rural % (95% CI)	Urban % (95% CI)	Total % (95% CI)
Hospital	73.8 (69.2-78.3)	92.8 (87.9-97.3)	75.9 (71.8-79.9)
Clinic	14.2 (9.9-18.5)	2.4 (0.1-4.6)	12.9 (9.1-16.7)
Private doctor	0.4 (0.0-1.0)	0.5 (0.0-1.5)	0.4 (0.0-0.9)
Home	10.7 (8.3-13.1)	3.9 (0.0-8.4)	10.0 (7.7-12.2)
Other	0.9 (0.0-1.8)	0.4 (0.0-1.2)	0.8 (0.0-1.6)

N=19309

$\chi^2_4 = 26.39$, p-value = 0.0001

Table 9.22 shows that amongst those who intend to use "Home" or "Other" as the place of delivery (Table 9.21), most (80.6%) currently pregnant women choose a relative for assisting them in delivery, with only 6.7% intending to use a village health care worker (VHW). The reasons for these respondents not intending to use health care facilities are shown in Table 9.23.

<u>Table 9.22</u> Distribution of assistance of delivery intended by currently pregnant women who plan to deliver at home

Assistance	Total % (95% CI)
VHW	6.7 (0.34-13.0)
Relative	80.6 (69.6-91.6)
Friend	2.4 (0.0-5.5)
Other	10.3 (1.8-18.8)

N=1828

A lack of money (38.1%) was the main reason cited by currently pregnant women for not intending to use health facilities for delivery. An error in the questionnaire design is that the responses "Birth came unexpected" and "Clinic was closed" were asked of currently pregnant women. At the time of the survey they would not have been able to respond to those questions. The 15.5% for "Birth came unexpected" therefore most probably refers to the pregnant women's previous delivery. The sample numbers for Tables 9.22 and 9.23 are too small to analyze by the urban/rural split.

Table 9.23 Distribution of reasons of currently pregnant women for not intending to use health facilities for delivery

Reasons	Total % (95% CI)
Lack of money	38.1 (27.1-49.1)
Birth came unexpected	15.5 (8.1-23.0)
Clinic too far	3.7 (0.0-7.9)
Clinic was closed	0.0
Prefer to deliver at home	12.2 (4.4-19.9)
Do not find it necessary	11.3 (3.3-19.2)
Attitude of clinic/hospital staff	7.6 (0.1-15.2)
Tradition is to deliver at home to learn to deliver others	0.0
Delivered at home to bury placenca	1.4 (0.0-4.1)
Other	10.2 (0.0-20.5)

N=1788

9.4.2.2 PAST PREGNANCIES

Table 9.24 shows that a significant association exists between place of residence and the place of delivery used by women who have given birth in the past. It should be noted that pregnant women who had not had a livebirth before the NH&NS was conducted, were not included in the results produced in this section. What is very noticeable from Table 9.24 is the high percentage of rural women who delivered at home (45.5) and the relatively low percentage who delivered at a hospital/clinic/ doctor (54.5%) compared to 88.1% who delivered at a hospital/ clinic/doctor in the urban areas. Also striking is the disparity between past experience and future intention. While 41% (Table 9.24) of last deliveries occurred at home, only 10% (Table 9.21) of currently pregnant women intended to deliver at home. This suggests that for approximately 31% of deliveries, the reasons for home delivery was beyond the woman's control. This is analysed further below (Table 9.26).

Table 9.24 **Distribution of the place of delivery of last baby (live births only).**

Place of delivery	Rural % (95% CI)	Urban % (95% CI)	Total % (95% CI)
Hospital/ Clinic/ Doctor	54.5 (51.4-57.6)	88.1 (84.6-91.5)	58.7 (55.9-61.5)
Home/Other	45.5 (42.3-48.6)	11.9 (8.5-15.3)	41.3 (38.5-44.1)
N	87156	12380	99537

$\chi^2_1 = 140.99$, p-value = 0.0000

A significant association exists between place of residence and the assistance of delivery used by women who have had live births.

Table 9.25 **Distribution of the assistance at delivery used by women who have delivered at home by place of residence (live births only)**

Assistance	Rural % (95% CI)	Urban % (95% CI)	Total % (95% CI)
VHW	8.5 (6.3-10.6)	6.0 (1.7-10.4)	8.4 (6.3-10.4)
Relative	76.5 (72.6-80.3)	63.3 (53.0-73.6)	76.0 (72.2-79.7)
Friend	9.5 (7.6-11.3)	18.6 (12.0-25.2)	9.8 (8.0-11.6)
Other	5.6 (3.3-7.8)	12.0 (2.9-21.2)	5.8 (3.6-8.0)
N	38050	1459	39510

$\chi^2_3 = 9.51$, p-value = 0.0292

Noticeable in Table 9.25 is the higher percentage of friends who assisted urban women (18.6%) than those who assisted rural women (9.5%). In contrast, a higher percentage of rural women had a relative assisting during a delivery (76.5%-rural versus 63.3%-urban). A likely explanation for these results is that in rural areas the extended family structure is more intact than in urban areas. There is the reasonable consistency between the findings in Tables 9.22 and 9.25.

A significant association exists between place of residence and the reasons for non-use of health-care facilities amongst women who have had livebirths (Table 9.26). A 'birth unexpected' is given as the main reason for not using a health care facility in both the rural (62.8%) and urban areas (69.6%). Another interesting result is that almost twice the proportion in rural areas (21.1%) as in urban areas (10.9%) give 'lack of money' as the reason for not using a health care facility. This may suggest that urban women have more money than their rural counterparts, and/or that the costs of getting to a facility (which, when in labour, will require motorized transport) is higher. At least 8.5% of women give reasons which indicate that it was their preference to deliver at home, which is consistent with the findings of Tables 9.21, 9.23 and 9.24.

In order to assess the determinants of past deliveries at a clinic or hospital by those women who have had at least one previous live birth (DELIVERY=1 indicates usage was made of a clinic or hospital for delivery, DELIVERY=0 indicates non-usage) a multiple logistic regression model was fitted in Table 9.27. The predictor variables fitted in this model were place of residence, age of mother, occupation level of mother, number of live births, education level and source of income.

Table 9.26 Distribution of the reasons for non-use of health-care facilities for last delivery, of women who were previously pregnant by place of residence (excluding currently pregnant)

Reasons	Rural % (95% CI)	Urban % (95% CI)	Total % (95% CI)
Lack of money	21.1 (18.1-24.2)	10.9 (5.7-16.1)	20.8 (17.8-23.7)
Birth came unexpectedly	62.8 (58.7-67.0)	69.6 (62.9-73.0)	63.1 (59.1-67.1)
Clinic was too far	2.4 (0.6-4.0)	4.0 (0.7-7.3)	2.4 (0.8-4.1)
Clinic was closed	0.2 (0.0-0.5)	0.0	0.2 (0.0-0.5)
Prefer to deliver at home	5.6 (4.1-7.1)	6.8 (1.8-11.8)	5.7 (4.2-7.2)
Do not find it necessary to deliver at clinic	2.4 (1.5-3.3)	1.7 (0.0-4.3)	2.4 (1.5-3.3)
Attitude of clinic/hospital staff	0.7 (0.2-1.2)	0.0	0.7 (0.2-1.1)
Tradition is to deliver at home to learn how to deliver others	0.2 (0.0-0.5)	0.0	0.2 (0.0-0.45)
Delivered at home so that could bury placenta	0.2 (0.0-0.4)	1.7 (0.0-3.8)	0.2 (0.0-0.5)
Other	4.2 (3.0-5.6)	5.3 (1.8-7.2)	4.3 (3.0-5.6)
N	37591	1440	39031

χ^2_9 = 20.88 , p-value = 0.0232

Table 9.27 Multiple Logistic Regression Model to predict past deliveries at a clinic or hospital (of past pregnancy)

DELIVERY	Degrees of freedom	S_waite Adj DF	S_waite Adj Chisq	p-value S_waite Adj Chisq
overall model	12	8.29	142.78	0.000
model minus intercept	11	8.91	152.25	0.000
rural / urban	1	1.00	68.96	0.000
education level	1	1.00	13.23	0.000
previous births	9	7.35	49.37	0.000

χ^2=40.0194, df=37, p-value=0.33762, ^{R2} = 0.100028

Education level: Educated - 'Standard 5 or above education level', Uneducated - 'below Standard 5 education level'.

Table 9.28 **Estimated Coefficients, Estimated Standard Errors, T-tests, and Odds Ratios for the Multiple Logistic Regression Model to determine the attendance or not of past pregnant women at a clinic/hospital for last delivery**

DELIVERY 1 = attended clinic/hospital 0 = did not attend	Beta Coeff	SE Beta	T-test Beta=0	P-value T-test Beta=0	Odds Ratio	Odds Ratio 95% CI
Constant	1.29	0.54	2.37	0.0201	3.63	(1.23;10.7)
Rural	-1.70	0.20	-8.30	0.0000	0.18	(0.12;0.27)
Urban	0.00	0.00	.	.	1.00	(1.00;1.00)
Uneducated	-0.44	0.12	-3.64	0.0005	0.65	(0.51;0.82)
Educated	0.00	0.00	.	.	1.00	(1.00;1.00)
Number of Births						
1	1.46	0.51	2.86	0.0054	4.32	(1.56;11.95)
2	0.77	0.51	1.51	0.1347	2.15	(0.78;5.90)
3	0.65	0.53	1.22	0.2276	1.92	(0.66;5.55)
4	0.81	0.54	1.51	0.1356	2.24	(0.77;6.50)
5	0.54	0.56	1.01	0.3147	1.72	(0.59;5.03)
6	0.28	0.55	0.51	0.6089	1.33	(0.44;4.01)
7	0.73	0.53	1.38	0.1714	2.08	(0.72;5.98)
8	0.70	0.56	1.24	0.2179	2.01	(0.66;6.13)
9	0.07	0.57	0.12	0.9046	1.07	(0.35;3.32)
10	0.00	0.00	.	.	1.00	(1.00;1.00)

Three effects (place of residence, education level, and the number of previous births) proved to be significant in Table 9.28 in predicting the attendance of ante-natal care. The effects age, occupation, and source of income were not significant and therefore not included in the model. Using Pearson's chi-square statistic there was not a significant lack-of-fit of the model (X^2=40.0194, 37 degrees of freedom, p-value=0.33762).

From the above logistic regression model rural woman are predicted to be only 0.18 times as likely as urban women to use a clinic/hospital for delivery. An odds ratio of 0.65 for 'uneducated' relative to 'educated' indicates that an uneducated pregnant woman is less likely than an educated pregnant woman to attend a clinic or hospital for delivery. Note that the only coefficient which is significantly different from zero is that of "0" previous birth, and the others are not monotonically decreasing. Therefore there does not appear to be a decreasing probability of attendance as the number of previous births increase beyond "0". This indicates that women are 2-4 times more likely to attend a clinic/hospital for their first births than for subsequent births. However the number of previous births beyond 0 do not affect the probability.

Of interest to the Lesotho Ministry of Health are those mothers who attend post-natal care (PNC) after delivery. From Table 9.29 it is seen that a high overall percentage of pregnant woman (88.6%) attend PNC. Significantly less mothers from rural areas attend PNC compared to mothers from urban areas (p-value=0.0012). An explanation for those mothers not attending PNC is shown in Table 9.30.

Table 9.29 Distribution of past pregnancies who received post natal care

	Rural % (95% CI)	Urban % (95% CI)	Total % (95% CI)
Received PNC	87.9 (85.8-90.0)	94.0 (92.0-95.9)	88.6 (86.7-90.5)
No PNC	11.4 (9.3-13.5)	5.56 (3.6-7.5)	10.7 (8.8-12.6)
Don't Know	0.7 (0.3-1.1)	0.47 (0.0-1.2)	0.7 (0.3-1.0)
N	83914	10101	94681

χ^2_2=14.62 , p-value=0.0012

A significant association exists between place of residence and the reasons for non-use of post-natal care facilities amongst women who have had livebirths (Table 9.30). The noticeable differences between rural and urban mothers regarding their reasons for not attending PNC are 'No money' (rural - 24.59%, urban - 11.6%) and 'Did not feel sick/mother and baby fine' (rural - 20.31%, urban - 9.9%) and 'No Reason' (rural - 21.3%, urban - 13.2%). 'Child passed away' is the main reason for mothers not attending PNC in the urban areas (30.8%) compared to only 10.4% in the rural areas.

Table 9.30 Distribution of reasons for not receiving post natal care after last pregnancy

	Rural % (95% CI)	Urban % (95% CI)	Total % (95% CI)
Did not feel sick/baby fine	20.3 (13.8-26.8)	9.9 (0.0-20.2)	19.6 (13.5-25.7)
No money	24.6 (19.3-29.9)	11.6 (3.3-19.8)	23.7 (18.8-28.7)
Child passed away	10.4 (7.1-13.8)	30.8 (18.7-43.0)	11.8 (8.4-15.1)
Received immunisation	4.6 (1.0-8.1)	7.0 (0.0-14.0)	4.7 (1.4-8.1)
Too young to visit clinic	10.1 (5.8-14.4)	13.2 (4.3-22.1)	10.3 (6.2-14.3)
Mother was ill	2.5 (0.9-4.1)	0.0	2.3 (0.8-3.8)
No reason	21.3 (13.8-28.7)	13.2 (2.1-24.3)	20.7 (13.7-27.7)
Other	6.2 (4.0-8.5)	14.3 (3.1-25.5)	6.8 (4.5-9.0)
N	8916	641	9557

χ^2_7=22.82 , p-value=0.0046

9.5 DISABILITY

Disability caused by injury or disease, whether incurred perinatally or later in life, must have a significant impact on the individual's quality of life and the economic prognosis for both the individual and the community as a whole.

The availability of good data on the cause and outcome of disability should yield important information which may lead to the planning of preventative programmes to reduce the incidence of this form of pathology and rehabilitation programmes to reduce the impact of disability. This would be expected to play an ever increasing role in the economy of developing countries such as Lesotho.

To enable the LMOH to gain a better understanding of disability in Lesotho, the various types of disabilities are related to their causes, by gender, age, and whether the respondents live in a rural or urban area. Finally a multiple logistic regression model is fitted to the data to predict the prevalence of disability from a number of demographic and occupational variables.

Table 9.31 Gender specific rates per 10,000 population for specific disabilities

DISABILITY	Female	Male	Total	p-value
Amput. both legs/feet	53.3 (41.[illegible]-65.1)	77.3 (57,7-96.9)	64.1 (50.4-77.8)	0.001
Lame/paral. both legs	29 [illegible] [illegible].2)	27.5 (17.7-37.3)	28.6 (20.8-36.4)	0.638
Amput of 1 foot/leg	47.6 (33.9-61.3)	71.5 (53.9-89.1)	58.3 (44.6-72.0)	0.000
Lame/paral. one leg	89.7 (68.1-111.3)	81.8 (70.0-93.6)	86.2 (70.5-101.9)	0.436
Amput. of hand(s)/arm (s)	36.4 (24.6-48.2)	52.8 (35.2-70.4)	43.7 (30.0-57.4)	0.007
Lame/paral. both arms	78.7 (59.1-98.3)	46.8 (37.9-56.6)	64.4 (50.7-78.1)	0.009
Lame / paral. 1 arm	27.2 (19.4-35.0)	36.9 (25.1-48.7)	31.5 (23.7-39.3)	0.042
Blind (Total)	23.6 (17.7-29.5)	27.4 (19.6-35.2)	25.3 (19.4-31.2)	0.320
Blind (partial)	88.6 (71.0-106.2)	100.6 (83.0-118.2)	94.0 (78.3-109.7)	0.127
Severe deafness	20.3 (14.4-26.2)	31.5 (23.7-39.3)	25.3 (19.4-31.2)	0.003
Speech problems	25.3 (45.2-33.1)	31.7 (23.9-39.)	28.2 (20.4-36.0)	0.146
Mentally ill/ retarded	11.9 (8.0-15.8)	14.4 (8.5-20.3)	13.0 (9.1-16.9)	0.000
Fits	2.3 (0.3-4.3)	5.2 (3.2-7.2)	3.6 (1.6-5.6)	0.042
Amput. of finger(s)	11.2 (7.3-15.1)	82.8 (71.0-94.6)	43.2 (37.3-49.1)	0.000
Amput. of toes	4.0 (2.0-6.0)	19.6 (13.7-25.5)	10.9 (8.9-12.9)	0.000
Other	19.0 (13.1-24.9)	27.0 (19.2-34.8)	22.6 (16.7-28.5)	0.037
Total	568.5	734.8	642.9	

For most disabilities the disability rate per 10000 population, shown in Table 9.31 is significantly higher amongst males than females especially with regard to digit and limb amputations. These disabilities include amputation of both legs/feet, amputation of hand(s)/arm(s), severe deafness, amputation of fingers and toes, and fits (only just significant and for a surprisingly low prevalence). The noticeable exception is with lameness/paralysis of both arms where females have almost double

Table 9.32 Disability rates in urban and rural areas of Lesotho

DISABILITY	Rural	Urban	p-value
Amput. both legs/feet	68.4 (52.7-84.1)	30.7 (17.0-44.4)	0.000
Lame/paral. both legs	29.8 (22.0-37.6)	19.0 (9.2-28.8)	0.110
Amput. of 1 foot/leg	61.6 (45.9-77.3)	32.5 (20.7-44.3)	0.004
Lame/paral. one leg	93.8 (76.2-111.4)	27.1 (13.4-40.8)	0.000
Amput. of hand/arm(s)	47.3 (31.6-63.0)	16.4 (8.6-24.2)	0.000
Lame/paral. both arms	68.2 (52.5-83.9)	35.6 (16.0-55.2)	0.011
Lame/paral. 1 arm	33.7 (23.9-43.5)	14.7 (6.9-22.5)	0.002
Blind (total)	26.7 (20.8-32.6)	14.4 (8.5-20.3)	0.005
Blind (partial)	99.8 (82.2-117.4)	49.0 (23.5-74.5)	0.002
Severe deafness	27.3 (21.4-33.2)	9.9 (4.0-15.8)	0.000
Speech problems	30.0 (22.2-37.8)	13.7 (5.9-21.5)	0.004
Mentally ill/ retarded	12.0 (8.1-15.9)	12.0 (4.2-19.8)	0.920
Fits	3.9 (1.9-5.9)	1.0 (0.0-3.0)	0.022
Amput. of finger(s)	46.9 (39.1-54.7)	14.9 (5.1-24.7)	0.000
Amput. of toes	11.9 (8.0-15.8)	4.1 (0.1-8.0)	0.002
Other	24.1 (18.2-30.0)	10.9 (5.0-16.8)	0.003
Total	686.3	307.6	

the rate of males.

Table 9.32 shows that there are significant differences between the rural and urban areas with regard to most disabilities (except for mental illness (p=0.920) and lame/paralysis in both legs (p=0.110)). The rural areas having significantly higher rates than the urban areas. There is an overall disability rate of 686.3 per 10000 in the rural areas as opposed to 307.6 per 10000 in the urban areas.

Table 9.33 shows that the disability rate for all disabilities increases with age. The only exception is mentally ill/retarded increases with age. The only exception is mentally ill/retarded where there is a (nonsignificant) decrease from 0-4 to 5-14 years.

The disabilities presented in Figure 9.5, are summarised into six main categories:

Upper limbs = Amputation of both legs/feet + Lame/paralysed in both legs + Lame/paralysed in one leg + Amputation of one foot/leg
Mental = Mentally ill/retarded
Lower limbs = Lame/paralysed in both arms + Amputation of hand(s)/arm(s) + Lame/paralysed in one arm
Digits = Amputation of finger(s) + Amputation of Toes
Speech/hearing = Speech problems + Severe deafness
Eyes = Blind (totally) + Blind (partially)

Certain disabilities are directly related to the ageing process - for example loss of hearing and sight. This is evident in the dramatic increase in these disability rates between age 45-59, and age 60+ (doubling in the case of blindness and quadrupling for deafness).

From Figure 9.5 it is also striking that the disabilities of leg(s)/foot(feet) (1060.2 per 10000) and hand(s)/arm(s) (688.3 per 10000) are very high in the 60 year and above age group. Since people with disabilities are also more likely to die earlier than those without, which would tend to decrease their prevalences in the older age groups. However, the data show that the "longer exposure" and age related causes, such as stroke, are a stronger influence on the prevalence of disability than the effect of death through disabilty.

Table 9.33 Age specific rates per 10,000 population for specific disabilities

	AGE GROUP						
DISABILITY	0-4	5-14	15-29	30-44	45-59	60+	p-value
Amput. both legs/feet	9.0 (3.1- 14.9)	18.7 (10.9- 26.5)	33.7 (21.9- 45.5)	86.5 (57.1- 115.9)	173.7 (126.7- 220.7	226.6 (169.8- 283.4)	0.000
Lame/paral. both legs	5.3 (1.4- 9.2)	12.0 (6.1- 17.9)	17.2 (9.4- 25.0)	29.1 (15.4- 42.8)	50.7 (29.1- 72.3)	124.5 (75.0- 174.0)	0.001
Amput. of 1 foot/leg	3.1 (0.0- 7.0)	15.0 (7.2- 22.8)	23.7 (13.9- 33.5)	86.8 (55.4- 118.2)	179.5 (130.5- 228.5)	202.0 (151.0- 253.0)	0.000
Lame/paral. one leg	6.7 (0.8- 12.6)	16.7 (10.8- 22.6)	22.7 (14.9- 30.5)	60.4 (36.9- 83.9)	204.9 (161.8- 248.0)	507.1 (405.0- 609.2)	0.000
Amput. of hand(s)/arm(s)	5.8 (0.0- 11.7)	18.2 (10.4- 26.0)	29.6 (15.9- 43.3)	59.5 (37.9- 81.1)	113.[illegible] (74.0- 152.4)	125.9 (78.9- 172.9)	0.000
Lame/paral. both arms	3.4 (0.0- 7.3)	11.4 (5.5- 17.3)	15.1 (9.2- 21.0)	41.1 (25.4- 56.8)	105.6 (70.3- 140.9)	449.9 (346.0- 553.8)	0.000
Lame/paral. 1 arm	3.2 (0.0- 7.1)	7.3 (3.4- 11.2)	14.1 (8.2- 20.0)	53.8 (34.2- 73.4)	83.8 (54.4- 113.2)	112.5 (76.7- 148.3)	0.000
Blind (total)	2.2 (0.0- 6.1)	6.9 (3.0- 10.8)	21.5 (13.7- 29.3)	42.8 (29.1- 56.5)	49.1 (29.5- 68.7)	83.0 (57.5- 108.5)	0.000
Blind (partial)	13.6 (5.8- 21.4)	23.1 (15.3- 30.9)	37.5 (25.7- 49.3)	89.8 (56.5- 123.1)	205.4 (162.3- 248.5)	482.4 (384.0- 580.8)	0.000
Severe deafness	0.0 .	7.0 (3.1- 10.9)	12.2 (6.3- 18.1)	18.1 (8.3- 27.9)	41.0 (25.3- 56.7)	152.8 (107.9- 197.7)	0.000
Speech problems	1.1 (0.0- 3.1)	7.7 (3.8- 11.6)	25.3 (13.5- 37.1)	40.5 (22.9- 58.1)	67.2 (43.7- 90.7)	86.9 (53.7- 33.2)	0.000
Mentally ill/ retarded	4.0 (0.0- 7.9)	2.0 (0.0- 5.9)	17.0 (11.1- 22.9)	16.0 (8.2- 23.8)	23.4 (11.6- 35.2)	27.6 (11.9- 43.3)	0.000
Fits	2.0 (0.0- 4.0)	2.0 (0.0- 4.0)	3.8 (1.8- 5.8)	6.7 (0.1- 12.6)	4.0 (0.0- 7.9)	6.5 (1.1- 12.4)	0.4[illegible]
Amput. of finger(s)	0.0 .	4.3 (2.3- 6.3)	15.9 (10.0- 21.8)	70.7 (51.1- 91.3)	124.9 (87.7- 162.1)	181.9 (148.7- 215.7)	0.000
Amput. of toes	1.1 (0.0- 3.1)	2.3 (0.0- 4.3)	3.3 (1.3- 5.3)	24.9 (11.2- 38.6)	24.9 (15.1- 38.6)	40.1 (24.3- 55.9)	0.000
Other	3.9 (0.0- 7.8)	8.0 (4.0- 12.0)	12.2 (6.3- 18.1)	32.4 (20.6- 44.2)	48.2 (20.8- 75.6)	85.3 (61.5- 109.1)	0.000
Total	65.3	163.6	306.2	760.9	1499.5	2919.6	

Figure 9.5 Age specific rates per 10 000 population of the most serious disabilities

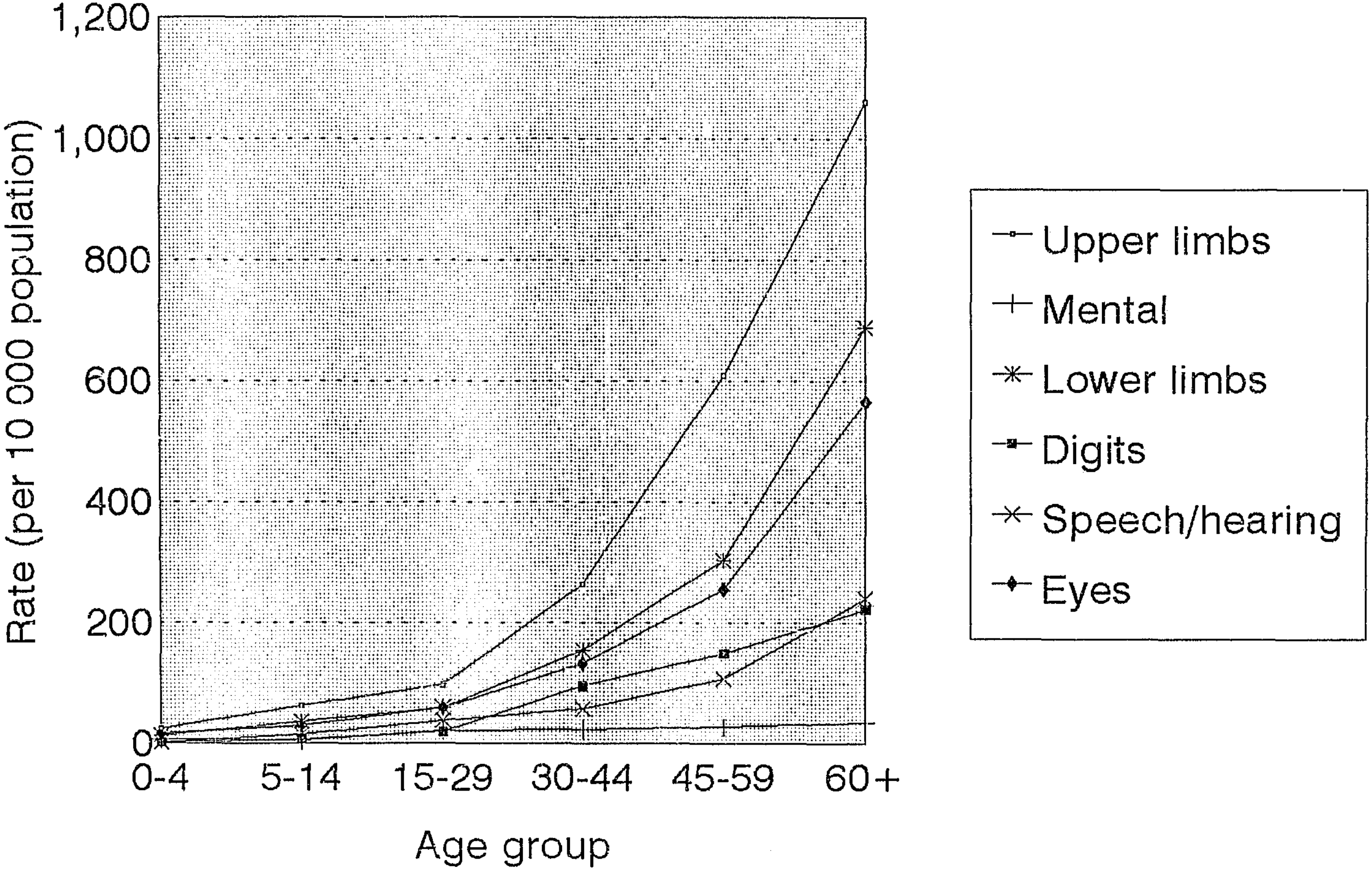


Table 9.34 shows the distribution of each type of disability recorded in the NH&NS by cause. Mining is identified by respondents as causing 17% of all disability with other occupations causing only 3.3% of disability. Noticeable are the high percentages of certain disabilities caused by mining: 20.8% - amputation of both legs/feet, 18.3% - amputation of hand/arm(s), 24.6% - amputation of one foot/leg, and 60.2% - amputation of fingers. The distribution of disabilities by cause is displayed in Figure 9.6. To investigate the full impact of mining on disability, only the disability amongst males over the age of 15 years who might be exposed to mining are shown in Table 9.35 and displayed in Figure 9.7. In this subgroup mining causes 36.9% of all disabilities, 70.8% of all finger amputations, 63.3% of all toe amputations and 33% of all deafness.

These results suggest that in the absence of mining there would be 36.9% less disabilities among adult males.

Table 9.34 Distribution of types of disabilities by cause (Percentage, (95% Confidence Interval))

	Born disabled	Illness	Traffic accident	Domestic accident	Mine accident	Other work/farm	Fight/ assault	Playing /sport	Horse	Witchcraft	Unknown	Other	Total no. in pop.
Amput. both legs/feet	3.9 (2.2-5.6)	14.4 (10.8-18.0)	5.1 (2.7-7.6)	13.2 (9.3-17.1)	20.8 (16.5-25.2)	3.8 (1.5-6.1)	5.0 (2.8-7.3)	4.8 (2.8-6.9)	3.7 (1.7-5.8)	5.1 (2.7-7.6)	10.8 (7.3-14.3)	9.2 (5.1-12.9)	8916
Lame/paral. both legs	13.2 (8.3-18.2)	35.9 (27.9-43.8)	2.6 (0.4-4.8)	0.8 (0.0-1.9)	6.9 (3.5-10.3)	0.8 (0.0-2.1)	2.4 (0.0-6.0)	0.7 (0.0-1.9)	0.0	1.6 (0.0-3.5)	28.4 (19.4-37.3)	6.7 (0.8-12.5)	3772
Amput. of one foot/leg	2.7 (0.8-4.6)	14.9 (9.1-20.6)	6.7 (3.9-9.5)	12.2 (7.7-16.7)	24.6 (19.3-30.0)	5.2 (2.4-8.0)	3.9 (2.4-5.4)	8.2 (5.5-10.8)	4.8 (2.4-7.1)	1.3 (0.2-2.3)	11.5 (7.5-15.5)	4.0 (1.0-7.0)	7222
Lame/paral. one leg	1.8 (0.0-2.9)	34.3 (26.8-41.8)	1.9 (0.5-3.3)	6.4 (3.8-9.1)	13.1 (9.0-17.2)	3.3 (1.3-5.2)	3.2 (1.6-4.8)	1.2 (0.2-2.3)	2.1 (0.6-3.[illegible])	1.2 (0.2-2.1)	21.6 (15.7-27.5)	10.0 (6.0-13.9)	9857
Amput. of hand/arm(s)	2.2 (0.6-3.7)	8.6 (5.0-12.2)	3.0 (0.8-5.1)	14.3 (8.9-19.7)	18.3 (14.1-22.6)	7.8 (4.2-11.3)	6.3 (3.3-9.2)	11.2 (7.4-14.8)	13.[illegible] (8.6-18.1)	1.6 (0.0-3.2)	7.1 (4.1-10.0)	6.4 (1.7-11.0)	5766
Lame/paral. both arms	3.2 (1.3-5.0)	47.6 (35.9-59.3)	0.7 (0.0-1.4)	1.8 (0.6-3.0)	3.8 (2.0-5.7)	1.8 (0.3-3.4)	2.1 (0.7-3.5)	0.8 (0.0-1.7)	0.3 (0.0-0.8)	0.9 (0.0-1.8)	25.6 (16.7-34.4)	11.5 (1.9-21.0)	8721
Lame/paral. one arm	3.1 (0.6-5.7)	12.4 (7.1-17.6)	5.2 (2.6-7.7)	15.5 (10.3-20.7)	12.6 (8.2-17.0)	2.1 (0.0-4.2)	8.6 (4.4-10.7)	9.9 (4.3-15.4)	10.2 (4.8-15.5)	2.6 (0.0-5.5)	16.3 (8.8-23.8)	1.5 (0.0-3.3)	3776
Blind (total)	12.9 (7.5-18.3)	29.3 (21.3-37.4)	1.4 (0.0-3.3)	4.1 (0.3-8.0)	5.2 (0.7-9.6)	0.0	3.2 (0.4-6.0)	0.0	0.0	10.7 (5.0-16.4)	24.7 (17.4-32.0)	8.4 (4.0-12.8)	3216
Blind partial	4.0 (1.8-6.1)	28.3 (22.0-34.7)	1.6 (0.7-2.5)	6.2 (3.8-8.6)	10.1 (8.0-12.2)	2.4 (0.5-4.2)	5.8 (3.6-8.1)	4.0 (2.4-5.7)	0.8 (0.0-1.6)	3.4 (1.9-4.9)	23.4 (17.7-29.1)	9.9 (6.2-13.6)	11697
Severe deafness	5.2 (1.6-8.7)	36.2 (25.7-46.7)	0.0	1.3 (0.0-3.2)	16.6 (10.4-22.7)	0.7 (0.0-2.5)	3.1 (0.4-5.9)	0.0	0.7 (0.0-2.0)	0.7 (0.0-2.0)	24.1 (15.2-32.9)	9.9 (6.2-13.6)	3363
Speech problems	5.2 (1.4-9.0)	18.7 (11.2-26.2)	7.4 (3.1-11.7)	3.7 (0.0-7.9)	13.8 (6.3-21.2)	4.1 (1.3-7.0)	7.1 (2.7-11.6)	1.4 (0.0-3.4)	2.8 (0.0-5.7)	6.0 (0.6-11.3)	21.4 (11.4-31.5)	8.3 (3.3-13.3)	3142
Mentally ill/retard	10.1 (0.0-20.6)	21.1 (10.1-32.2)	2.0 (0.0-5.7)	0.8 (0.0-2.4)	5.1 (0.0-10.3)	3.4 (0.0-7.9)	1.7 (0.0-5.1)	0.0	0.9 (0.0-2.7)	12.7 (4.7-20.6)	31.1 (22.2-40.1)	10.5 (3.6-17.4)	1538

	Born disabled	Illness	Traffic accident	Domestic accident	Mine accident	Other work/farm	Fight/ assault	Playing /sport	Horse	Witchcraft	Unknown	Other	Total no. in pop.
Fits	9.7 (0.0- 21.3)	23.4 (0.0- 47.2)	0.0	16.3 (0.0- 33.9)	4.9 (0.0- 14.3)	0.0	0.0	0.0	0.0	5.2 (0.0- 15.3)	32.5 (3.5- 61.5)	8.1 (0.0- 20.2)	457
Amput. of Finger(s)	1.6 (0.0- 3.2)	4.9 (2.3- 7.6)	1.5 (0.1- 2.9)	12.4 (8.5- 16.3)	60.2 (53.3- 67.2)	5.8 (2.7- 8.8)	3.9 (1.3- 6.5)	1.6 (0.2- 3.1)	2.1 (0.3- 3.8)	0.4 (0.0- 1.3)	2.5 (0.3- 4.6)	3.1 (0.4- 5.7)	5340
Amput. of Toes	2.3 (0.0- 6.2)	18.0 (7.0- 29.0)	3.6 (0.0- 8.8)	5.2 (0.0- 10.7)	49.7 (38.0- 61.3)	3.7 (0.0- 8.5)	0.0	8.0 (1.4- 14.5)	2.8 (0.0- 6.6)	1.3 (0.0- 3.7)	4.4 (0.0- 9.6)	1.0 (0.0- 3.0)	1272
Other	4.4 (1.0- 7.8)	25.1 (16.0- 34.2)	3.7 (0.4- 7.0)	7.5 (3.4- 11.7)	15.1 (7.3- 22.8)	2.8 (0.1- 5	8.6 (3.7- 13.5)	5.7 (2.0- 9.4)	2.7 (0.1- 5.3)	1.9 (0.0- 4.5)	13.3 (7.5- 19.1)	9.1 (4.1- 14.2)	2851
Total	4.2 (3.3- 5.1)	24.61 (20.6- 28.6)	2.9 (2.3- 6.9)	0.8 (0.0- 2.4)	17.0 (15.3- 18.7)	3.3 (2.4- 4.2)	4.5 (3.7- 5.2)	3.9 (3.1- 4.7)	3.1 (2.4- 3.8)	2.8 (2.1- 3.5)	17.9 (14.5- 21.3)	8.0 (5.5- 10.6)	80906

χ^2_{80}=32437, p-value=0.0000

FIGURE 9.6 TYPE OF DISABILITY BY CAUSE

FOR THE WHOLE POPULATION - LESOTHO 1989

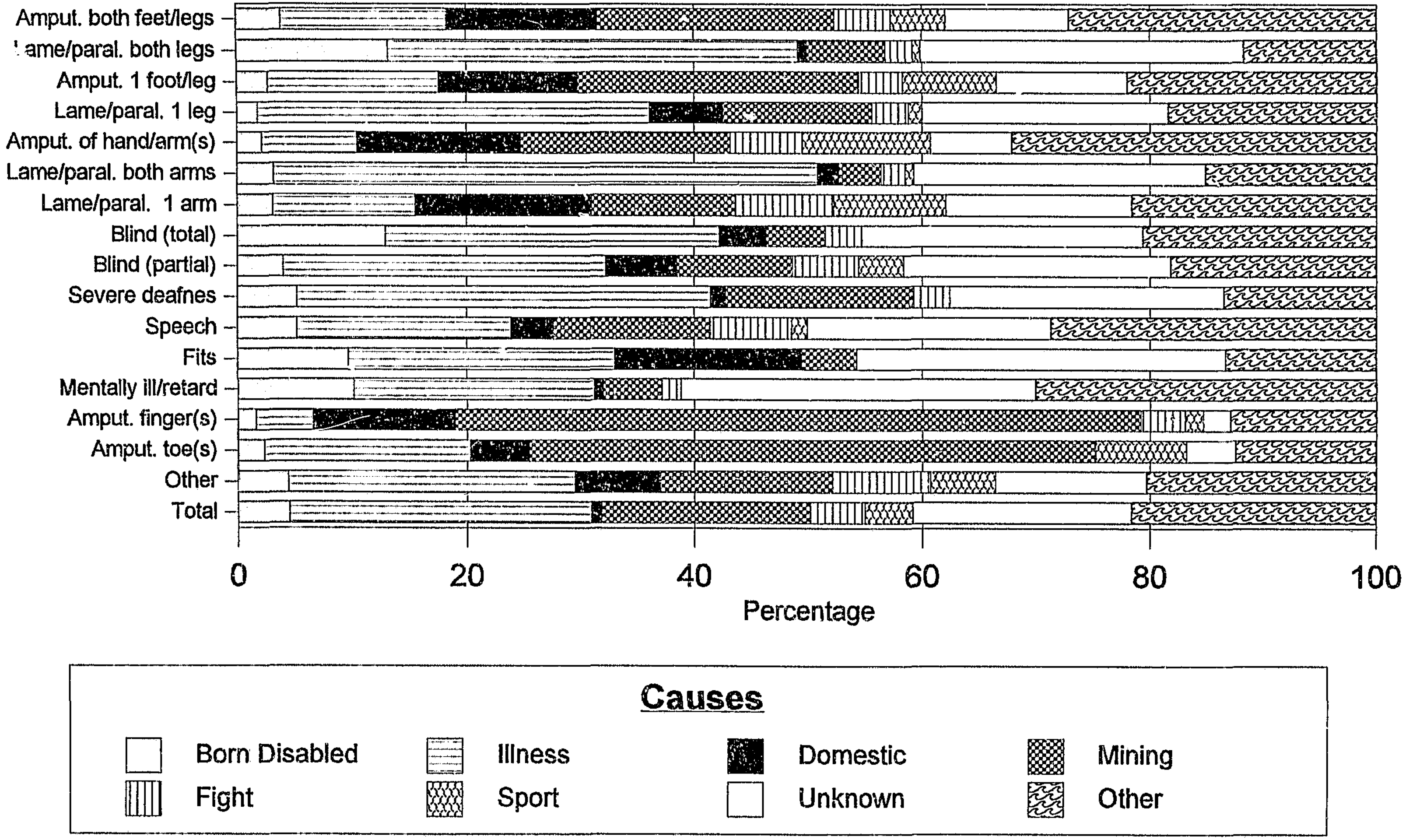


Table 9.35 Distribution of types of disabilities by cause (males older than 15 years)

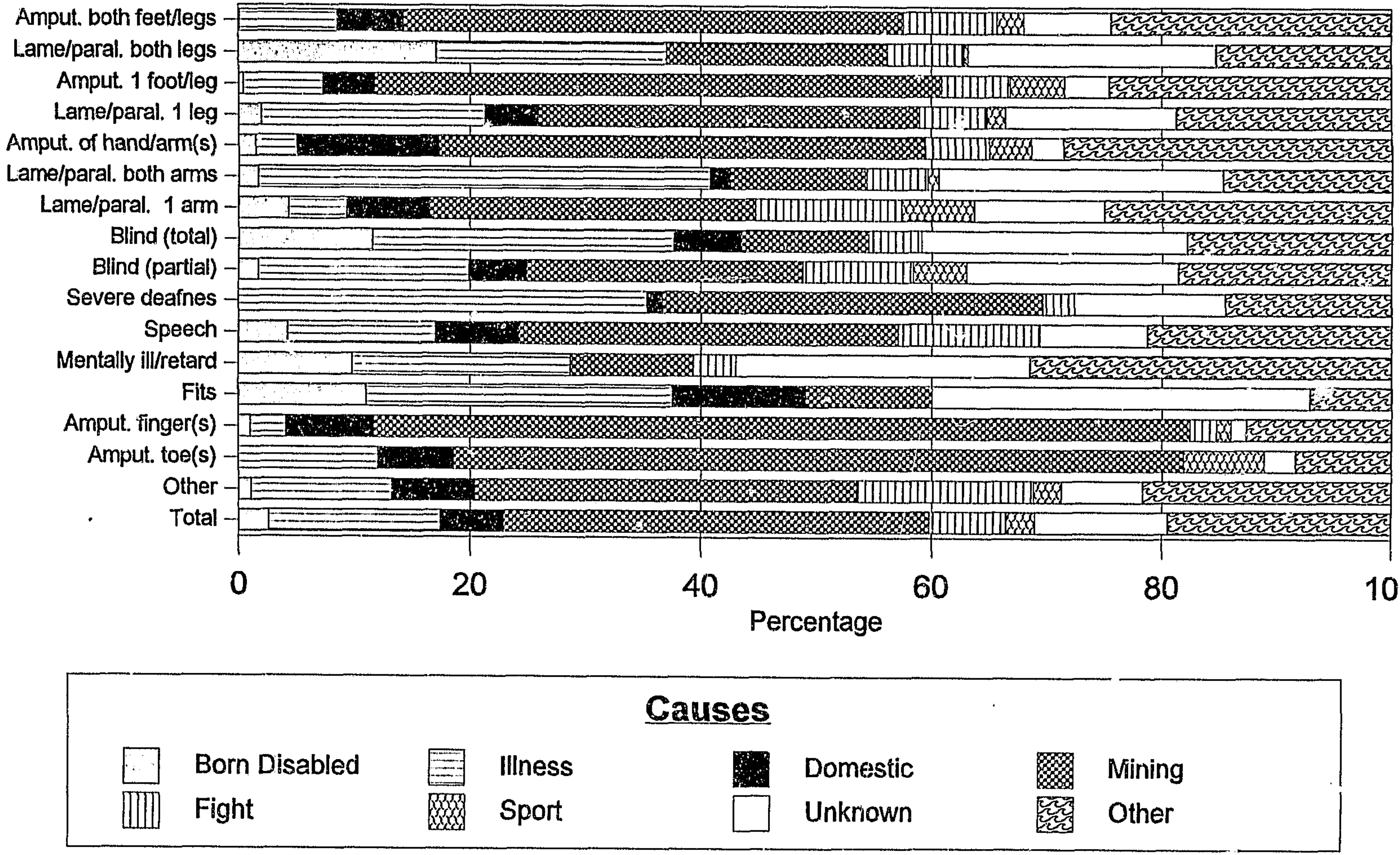
	Born disabled	Illness	Traffic accident	Domestic accident	Mine accident	Other work/farm	Fight/ assault	Playing/ sport	Horse	Witchcraft	Unknown	Other	Total no.& % in pop.
Amput. both legs/feet	0.1 (0.0-0.4)	8.4 (4.9-11.9)	7.5 (2.6-12.4)	5.7 (3.0-8.4)	43.4 (36.0-50.7)	3.5 (0.5-6.1)	8.1 (4.7-11.5)	2.3 (0.3-4.4)	4.4 (1.3-7.5)	5.4 (1.1-9.7)	7.6 (3.5-11.7)	3.4 (0.9-5.9)	4277 11.5%
Lame/paral. both legs	17.1 (7.7-26.5)	20.0 (11.7-28.3)	4.6 (0.0-9.7)	0.0 .	19.1 (10.3-27.9)	2.2 (0.0-5.8)	6.6 (0.0-16.4)	0.4 (0.0-1.1)	0.0 .	1.7 (0.0-5.0)	21.5 (10.4-32.6)	6.7 (0.0-14.3)	1363 3.6%
Amput. of one foot/leg	0.4 (0.0-1.3)	6.8 (2.4-11.3)	7.5 (3.5-11.5)	4.5 (2.0-7.0)	49.2 (40.4-57.9)	4.8 (2.1-7.5)	5.9 (3.0-8.8)	4.7 (1.8-7.6)	8.2 (3.6-12.7)	2.0 (0.0-4.0)	3.9 (0.6-7.2)	2.0 (0.0-4.3)	3562 9.6%
Lame/paral. one leg	2.0 (0.0-4.3)	19.3 (11.7-27.0)	2.2 (0.2-4.2)	4.6 (0.9-8.3)	33.1 (24.4-41.8)	4.2 (0.4-8.0)	5.8 (2.5-9.1)	1.6 (0.0-3.4)	3.3 (0.3-6.3)	0.6 (0.0-1.7)	14.9 (8.9-20.9)	8.3 (4.4-12.1)	3963 10.7%
Amput. of hand/arm(s)	1.5 (0.0-3.8)	3.5 (0.2-6.8)	1.3 (0.0-2.8)	12.3 (4.8-19.7)	42.1 (34.5-50.1)	7.7 (2.4-13.0)	5.6 (1.2-10.0)	3.6 (0.2-7.0)	14.5 (7.6-21.3)	1.0 (0.0-3.0)	2.8 (0.0-6.7)	4.2 (0.0-9.4)	2567 6.9%
Lame/paral. both arms	1.7 (0.0-4.1)	39.1 (28.2-50.0)	0.8 (0.0-2.4)	1.7 (0.0-4.0)	11.9 (5.9-17.8)	1.9 (0.0-4.3)	5.3 (1.2-9.4)	0.9 (0.0-2.7)	0.0 .	0.9 (0.0-2.5)	24.7 (17.1-32.2)	11.1 (1.9-20.3)	2679 7.2%
Lame/paral. one arm	4.3 (0.0-9.1)	5.0 (0.0-10.1)	4.5 (0.0-9.1)	7.2 (0.5-13.9)	28.2 (17.8-38.7)	3.3 (0.0-7.2)	12.7 (4.1-20.7)	6.3 (0.4-12.1)	16.4 (7.1-25.7)	0.0 .	11.3 (3.2-19.4)	0.7 (0.0-2.2)	1684 4.5%
Blind (total)	11.5 (4.2-18.9)	26.1 (14.8-37.4)	3.0 (0.0-7.1)	5.8 (0.0-11.5)	11.1 (2.0-20.2)	0.0 .	4.6 (0.0-9.7)	0.0 .	0.0 .	11.0 (3.2-18.8)	23.1 (13.2-32.9)	3.8 (0.0-8.5)	1502 4.0%
Blind partial	1.7 (0.0-3.4)	18.2 (11.6-24.8)	1.4 (0.0-3.0)	5.0 (1.6-8.4)	23.9 (19.4-28.4)	4.7 (0.9-8.5)	9.6 (4.8-14.4)	4.6 (2.1-7.0)	0.9 (0.0-2.2)	4.0 (1.4-6.6)	18.4 (11.4-25.4)	7.6 (3.4-11.8)	4856 13.1%
Severe deafness	0.0 .	35.3 (23.1-47.5)	0.0 .	1.3 (0.0-3.9)	33.0 (22.6-43.4)	1.3 (0.0-4.0)	2.7 (0.0-6.5)	0.0 .	1.3 (0.0-4.0)	1.3 (0.0-4.0)	13.2 (4.9-21.5)	10.4 (1.1-19.7)	1687 4.5%
Speech problems	4.2 (0.0-8.9)	12.7 (3.7-21.8)	7.9 (1.2-14.6)	7.2 (0.0-16.6)	32.9 (17.8-47.9)	2.2 (0.0-5.7)	12.2 (4.1-20.3)	0.0 .	3.3 (0.0-8.1)	3.4 (0.0-7.9)	9.4 (1.2-17.6)	4.5 (0.0-9.7)	1315 3.5%
Mentally ill/retard	9.8 (0.0-22.5)	18.9 (3.7-34.1)	0.0 .	0.0 .	10.6 (0.0-21.3)	3.0 (0.0-8.8)	3.7 (0.0-9.7)	0.0 .	1.9 (0.0-5.5)	14.6 (2.1-27.0)	25.5 (9.9-41.1)	12.1 (2.2-22.0)	742 1.9%

	Born disabled	Illness	Traffic accident	Dcmestic accident	Mine accident	Other work/farm	Fight/ assault	Playing/ sport	Horse	Witchcraft	Unknown	Other	Total no.& % in pop.
Fits	11.0 (0.0- 31.2)	26.5 (0.0- 57.2)	0.0 .	11.5 (0.0- 32.6)	11.0 (0.0- 31.4)	0.0 .	0.0 .	0.0 .	0.0 .	0.0 .	32.9 (0.0- 73.1)	14.3 (0.0- 30.8)	201 0.5%
Amput. of Finger(s)	1.0 (0.0- 2.5)	3.0 (0.6- 5.4)	1.8 (0.2- 3.4)	7.6 (3.9- 11.3)	70.8 (63.2- 78.4)	6.2 (2.6- 11.8)	2.3 (0.2- 4.4)	1.2 (0.0- 2.6)	2.0 (0.1- 3.9)	0.5 (0.0- 1.6)	1.4 (0.0- 3.0)	2.1 (0.0- 4.3)	4473 12.0%
Amput. of Toes	0.0 .	12.0 (1.9- 22.1)	4.6 (0.0- 11.2)	6.6 (0.0- 13.6)	63.3 (51.2- 75.4)	0.0 .	0.0 .	7.0 (0.0- 14.4)	3.6 (0.0- 8.4)	0.0 .	2.8 (0.0- 7.6)	0.0 .	998 2.7%
Other	1.1 (0.0- 3.3)	12.1 (4.7- 19.6)	2.7 (0.0- 6.7)	7.2 (0.5- 13.9)	33.1 (19.2- 47.0)	4.0 (0.0- 8.7)	15.3 (5.1- 25.5)	2.4 (0.0- 6.1)	0.4 (0.0- 1.3)	2.4 (0.0- 7.1)	7.1 (0.8- 13.4)	12.2 (3.1- 21.3)	1300 3.5%
Total	2.6 (1.7- 3.5)	14.8 (11.9- 17.7)	3.4 (2.3- 4.5)	5.5 (4.2- 6.8)	36.9 (34.0- 39.8)	3.9 (2.7- 5.1)	6.6 (5.4- 7.8)	2.5 (1.8- 3.2)	4.1 (2.8- 5.4)	2.6 (1.8- 3.4)	11.6 (9.2- 14.0)	5.6 (3.7- 7.5)	37169

χ^2_{80}=32437, p-value=0.0000

FIGURE 9.7 TYPE OF DISABILITY BY CAUSE

FOR MALES > 15 YEARS - LESOTHO 1989



The results of a multiple logistic regression model fitted to the data to predict the prevalence of disability are given in Table 9.36. Six effects, place of residence, age, sex, education level, mining, and occupation level and the interactions age X mining and sex X age proved to be significant in predicting the prevalence of disability. The source of income variable was not significant and therefore is not included in the model. Even though a lack-of-fit of this model is indicated ($\hat{C} = 44.989$, $p < 0.0001$), the significant effects may still be interpreted. Below are possible reasons for the lack-of-fit:

- The predictor variables that could have added to a better fitting model were not included in the survey.
- Disability is very specific to an individual and his/her particular circumstances and therefore cannot always be related to specific variables or combinations thereof.
- Disability includes many different specific disabilities each likely to have its own predicted model. When these disabilities are combined, except for a few effects which are common to all, it explains only a small part of the variation.

Disability occurs 1.871 times more frequently amongst rural respondents than among urban respondents. The occurrence of disability is less frequent among females than among males (0.867 times) although not significantly so. For every increase of one year in age the risk of disability increases 1.045 times. Therefore for every increase of 20 years in age, the risk of disability increases 2.411 times (from Table 9.36, $\exp(20\text{x}0.044) \approx 2.411$). Disability occurs 1.372 times more for uneducated respondents than for educated ones.

Table 9.36 Estimated Coefficients, Estimated Standard Errors, T-tests, and Odds Ratios for the Multiple Logistic Regression Model to predict the presence or absence of disability

DISABILITY 1 = has a disability 0 = has no disability	Beta Coeff	SE Beta	T-test Beta=0	P-value T-test Beta=0	Odds Ratio	Odds Ratio 95% CI
Constant	-5.125	0.188	-27.26	0.0000	0.006	(0.004;0.009)
Rural	0.626	0.160	3.92	0.0002	1.871	(1.361;2.571)
Urban	0.000	0.000	.	.	1.000	(1.000;1.000)
Sex						
1	-0.142	0.124	-1.15	0.2528	0.867	(0.678;1.109)
2	0.000	0.000	.	.	1.000	(1.000;1.000)
Age	0.044	0.002	17.52	0.000	1.045	(1.040;1.051)
Uneducated	0.316	0.064	4.94	0.000	1.372	(1.208;1.558)
Educated	0.000	0.000	.	.	1.000	(1.000;1.000)
Mining						
1	1.268	0.223	5.68	0.0000	3.554	(2.279;5.541)
2	1.192	0.237	5.03	0.0000	3.293	(2.055;5.276)
3	0.647	0.302	2.14	0.0355	1.910	(1.046;3.4870
4	0.000	0.000	.	.	1.000	(1.000;1.000)
Occupation						
1	0.413	0.257	1.60	0.1126	1.511	(0.905;2.522)
2	0.230	0.106	2.17	0.0332	1.258	(1.019;1.554)
3	0.422	0.132	3.20	0.0020	1.526	(1.173;1.984)
4	0.388	0.119	3.25	0.0017	1.475	(1.163;1.870)
5	0.528	0.114	4.63	0.0000	1.695	(1.351;2.127)
6	-0.198	0.135	-1.47	0.1453	0.820	(0.628;1.072)
7	0.231	0.128	1.80	0.0752	1.260	(0.976;1.625)
8	0.913	0.165	5.52	0.0000	2.492	(1.792;3.463)
9	1.066	0.154	6.90	0.0000	2.904	(2.136;3.948)
10	0.000	0.000	.	.	1.000	(1.000;1.000)
Sex X age						
1	-0.006	0.003	-2.02	0.0465	0.994	(0.988;1.000)
2	0.000	0.000	.	.	1.000	(1.000;1.000)
Mining X age						
1	-0.020	0.004	-4.97	0.0000	0.980	(0.972;0.988)
2	-0.019	0.005	-3.77	0.0003	0.981	(0.972;0.991)
3	-0.013	0.006	-2.06	0.0426	0.987	(0.975;1.000)
4	0.000	0.000	.	.	1.000	(1.000;1.000)

$R^2 = 0.111$, $\hat{C} = 44.989$, $p < 0.0001$

Sex: "Female" = 1, "Male" = 2

Education: "Educated" - Standard 5 or above education level, "Uneducated" - below Standard 5 education level.

Occupation: "Employer" = 1, "Own account worker/farmer" = 2, "Unpaid family worker" = 3, "Casual worker" = 4, "Unemployed" = 5, "Student" = 6, "Homemaker" = 7, "Pensioner/retired" = 8, "Other" = 9, "Regular wage/salary earner" = 10

Mining: "> 10 years" = 1, "5 - 10 years" = 2, "Less than 5 years" = 3, "Never underground" = 4,

With regular wage/salary earners regarded as the reference group for occupation, and consicering only the significant effects, own account workers or farmers are 1.258 times more frequently disabled, unpaid family workers are 1.526 times more frequently disabled, casual workers 1.475 times more frequent, unemployed respondents 1.69 times more frequent, students 0.82 times less frequent, a pensioner/retired respondent 2.492 times more frequent and other respondents whose occupation was not classified, 2.904 times more frequent. The effect of mining indicates that the longer a person works underground the more likely he is to be disabled. Relative to 'not working underground' respondents with 'more than 10 years underground' are 3.554 times more frequently disabled, '5-10 years underground' are 3.293 times more frequently disabled and 'less than 5 years underground' 1.91 times more frequently disabled. The sex X age interaction indicates that the age effect is less strong in females than in males. The mining X age effect is negative within each of the "length of mining" categories while overall it is positive. This is because age appears implicitly in the length of mining categories, so there is an element of "double counting" of age in the main effects: age and mining. The interaction effect tends to correct for this double counting.

9.6 INJURIES

As in the case of disability, injuries can have a significant impact on the quality of life of an individual and adversely affect both the individual and community economically. In this section the extent of injuries in Lesotho at the time of the NH&NS are assessed by relating them to age, gender, place of residence and the cause of injury. The recall period prior to the survey was considered during which the injury might have occurred was, on average, 11 days. The period was based on injuries since the preceding Sunday. Therefore the interval could vary from 8 to 13 days. Note that injuries that result in hospitalisation for more than this period would not have been detected by the survey - however the rate of such severe injuries is very low.

Table 9.37 Gender specific rates of injuries per 10,000 population with 95% confidence limits (based on an 11 day period prior to the survey)

INJURY	Female	Male	Total	p-value
Broken bone	13.0 (9.1-16.9)	22.0 (14.2-29.8)	17.0 (13.1-20.9)	0.0081
Torn skin or flesh	87.0 (69.4-104.6)	97.0 (75.4-118.6)	91.0 (73.4-108.6)	0.2408
Burns	19.0 (15.1-22.9)	16.0 (10.1-21.9)	17.0 (13.1-20.9)	0.3705
Head injury	7.0 (3.1-10.9)	17.0 (11.1-22.9)	12.0 (8.1-15.9)	0.0000
Sprain or bruise	29.0 (19.2-34.9)	33.0 (23.2-42.8)	31.0 (23.2-38.8)	0.4316
Eye injury	3.0 (1.0-5.0)	5.0 (3.0-7.0)	4.0 (2.0-6.0)	0.0656
Other	11.0 (7.1-14.9)	14.0 (8.1-19.9)	12.0 (8.1-15.9)	0.3078

χ^2_7=30.95, p-value=0.0003

A significant overall association exists between gender and type of injury (p=0.0003) in Table 9.37. When comparing specific injury rates of males and females, it is clear that males have a significantly higher rate of broken bones and head injuries than females. In Figure 9.9 the cause of these injuries by gender are shown.

Table 9.38 Injury rates per 10 000 population in urban and rural areas of Lesotho (based on an 11 day interval period prior to the survey)

INJURY	Rural	Urban	p-value
Broken bone	17.0 (11.1-22.9)	17.0 (9.2-24.8)	0.9490
Torn skin or flesh	92.0 (72.4-111.6)	88.0 (52.7-123.2)	0.8641
Burns	17.0 (13.1-20.9)	23.0 (11.2-34.8)	0.2467
Head injury	12.0 (8.1-15.9)	7.0 (1.1-12.9)	0.1026
Sprain or bruise	31.0 (21.2-40.8)	27.0 (17.2-36.8)	0.5452
Eye injury	4.0 (2.0-6.0)	5.0 (1.1-8.9)	0.5280
Other	13.0 (7.1-18.9)	8.0 (2.1-13.9)	0.1280

$\chi^2_7=9.21$, p-value=0.2536

From Table 9.38 it is clear that there is no significant difference between the rural and urban areas in any of the injury rates.

From Table 9.39, with the exception of burns, eye injuries and "other" injuries the injury rate increases from the 0-4 year age category to the 50-59 age category, and then decreases for the above 60 years age category. There is a very strong overall association between age group and injury rate as well as there being a significant difference between the age groups when considering each injury separately. Burns are 3 times more common in the 0-4 age group than in older groups.

Table 9.39 Age specific injury rates per 10 000 population in Lesotho, 1989 (based on an 11 day period prior to the survey)

INJURY	0 - 4	5 - 14	15 - 29	30 -44	45 - 59	60+	p-value
Broken bone	5.0 (1.1-8.9)	9.0 (5.1-12.9)	20.0 (10.2-29.8)	20.0 (12.2-27.8)	35.0 (19.3-50.7)	29.0 (13.3-44.7)	0.0007
Torn skin or flesh	76.0 (50.5-101.5)	89.0 (67.4-110.6)	80.0 (62.4-105.6)	119.0 (87.6-150.4)	114.0 (84.6-143.4)	73.0 (43.6-102.4)	0.0096
Burns	41.0 (29.2-52.8)	12.0 (6.1-17.9)	14.0 (8.1-19.9)	17.0 (9.2-24.8)	16.0 (6.2-25.8)	5.0 (0.0-10.9)	0.0000
Head injury	5.0 (1.0-8.9)	13.0 (7.1-18.9)	16.0 (10.1-21.9)	16.0 (6.2-25.8)	13.0 (1.2-24.8)	1.0 (0.0-3.0)	0.0006
Sprain or bruise	11.0 (3.2-18.8)	20.0 (10.2-29.8)	30.0 (20.2-39.8)	44.0 (26.4-61.6)	61.0 (37.5-84.5)	45.0 (21.5-68.5)	0.0004
Eye injury	1.0 (0.0-3.0)	4.0 (0.0-7.9)	5.0 (1.1-8.9)	4.0 (0.0-7.9)	0.0	8.0 (0.0-15.8)	0.0019
Other	3.0 (0.0-6.9)	12.0 (4.1-19.8)	13.0 (5.2-20.8)	15.0 (7.2-22.8)	10.0 (2.2-17.8)	28.0 (10.4-45.6)	0.0093

χ^2_{35}=196.66, p-value=0.0000

An overall significant association exists between injury and cause of injury (p=0.000). Domestic accidents (33.9%) and playing/sport (23.9%) account for more than 50% of all the injuries (see Figure 9.8). Noticeable is the high percentage of burns (60%) caused by domestic accidents, the high percentage of head injuries caused by fights/assaults (39.4%), and the high percentage of eye injuries that are work/farm (34.4%) related. In Table 9.37 it was shown that significant gender differences exist for the injuries 'broken bones' and 'head injuries'. In Figure 9.9 the causes of the injuries responsible for the gender differences are shown.

FIGURE 9.8 TYPE OF INJURY BY CAUSE

FOR WHOLE POPULATION - LESOTHO 1989

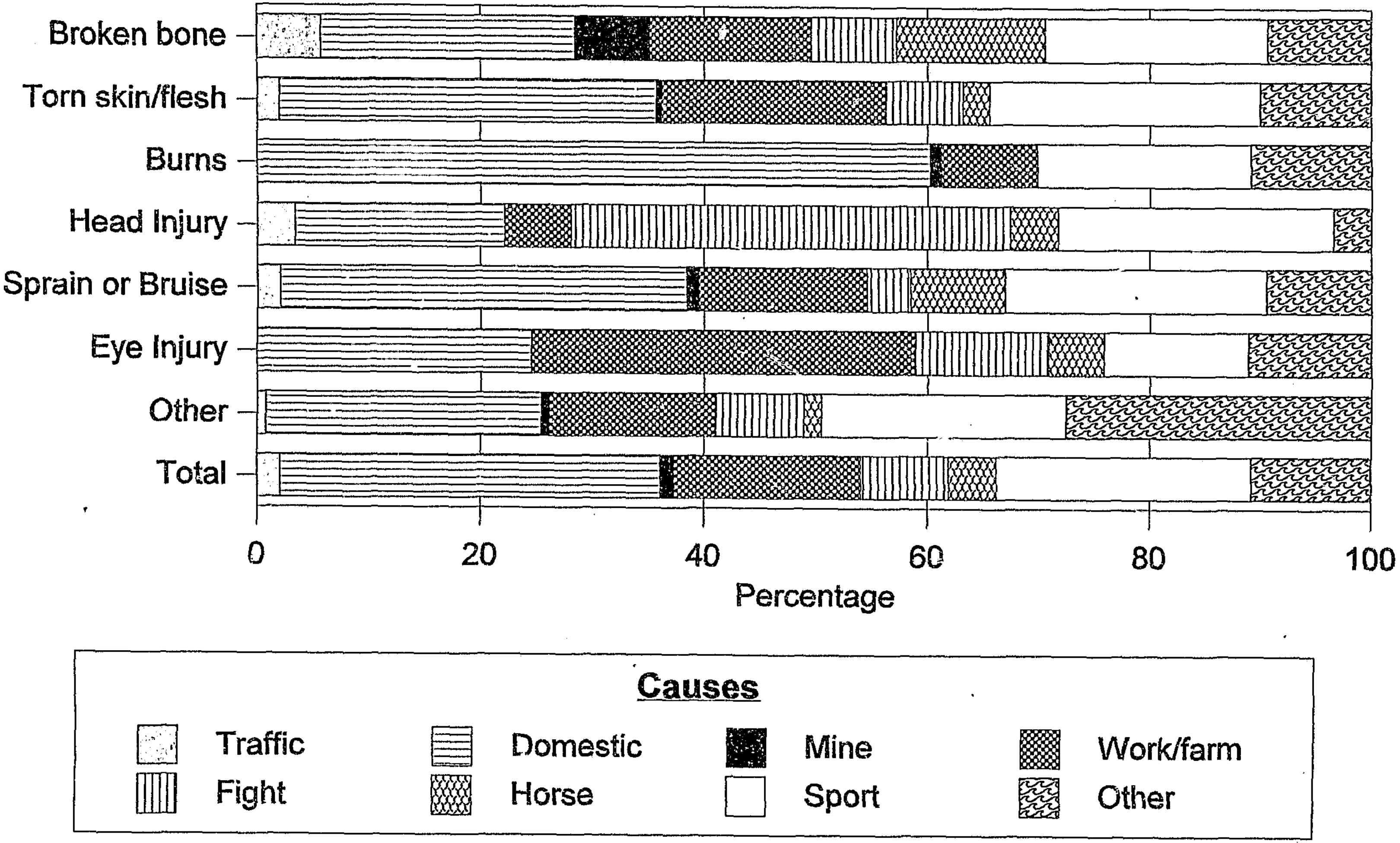


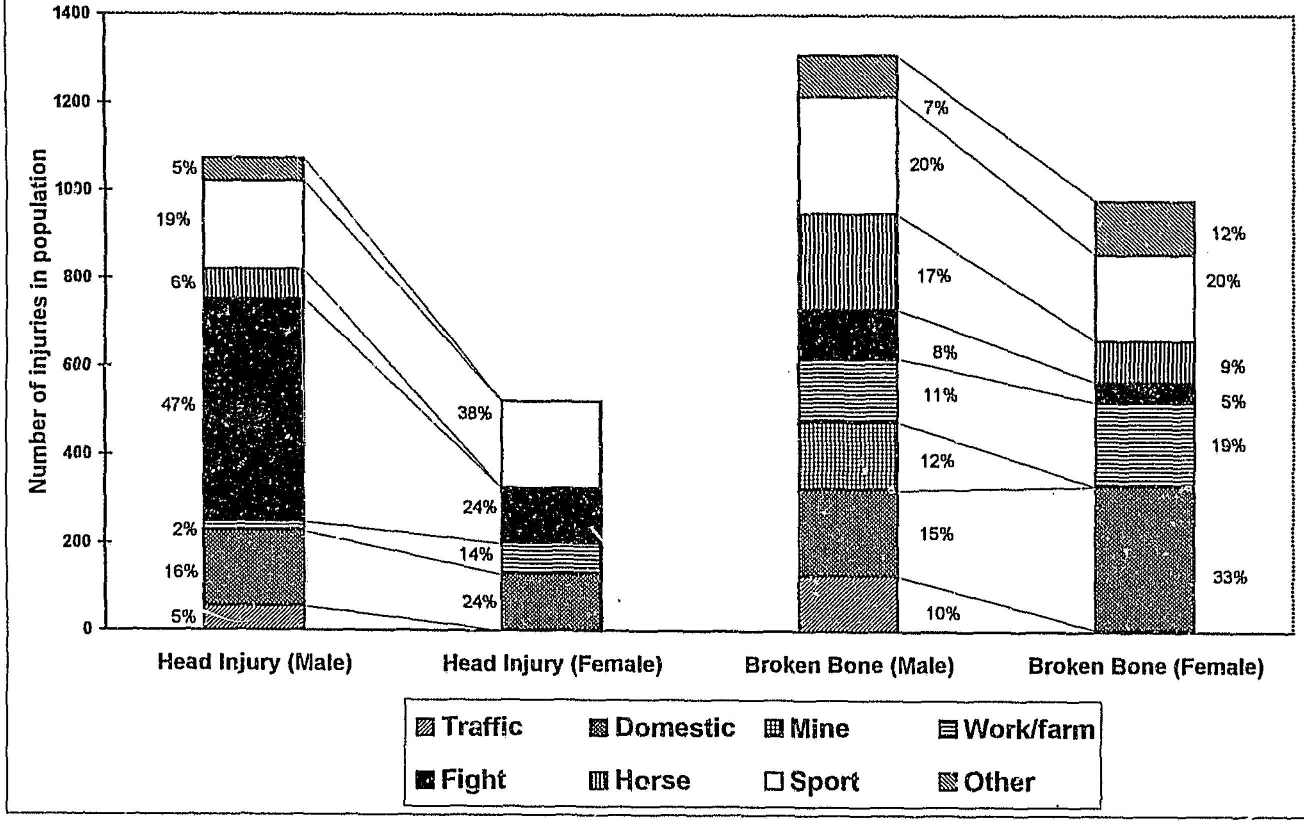
Table 9.40 Distribution of injury by cause based on an 11 day period prior to the survey

INJURY	Traffic	Domestic accidents	Mine	Work/ Farm	Fight/ Assault	Horse	Playing, sport	Other	Total no. & % in pop.
Broken bone	5.8 (2.1-9.5)	22.9 (15.9-29.8)	6.8 (0.5-13.1)	14.6 (5.9-23.3)	7.7 (2.5-12.8)	13.5 (4.4-22.6)	20.3 (13.3-27.2)	9.4 (3.7-15.1)	2111 9.1%
Torn skin or flesh	2.0 (0.5-3.5)	33.7 (26.4-41.0)	0.5 (0.0-1.1)	20.1 (13.8-26.4)	6.9 (4.9-8.8)	2.4 (1.2-3.5)	24.4 (20.3-28.5)	10.0 (6.5-13.9)	11597 50.1%
Burns	0.0	60.2 (49.8-70.6)	1.0 (0.0-3.0)	8.6 (2.5-14.7)	0.0	0.0	19.3 (12.4-26.3)	10.8 (4.6-16.9)	2047 8.8%
Head injury	3.4 (0.0-7.6)	18.8 (8.6-28.9)	0.0	5.9 (0.0-12.0)	39.4 (25.0-53.9)	4.3 (0.0-8.9)	24.9 (15.0-34.7)	3.4 (0.0-7.9)	1437 6.2%
Sprain or bruise	2.1 (0.1-4.1)	36.3 (26.2-46.3)	1.1 (0.0-2.6)	15.0 (8.5-21.5)	3.9 (1.6-6.2)	8.5 (4.2-12.8)	23.6 (17.0-30.2)	9.4 (4.3-14.5)	3985 17.2%
Eye injury	0.0	24.5 (8.9-40.1)	0.0	34.4 (14.8-54.0)	11.9 (0.9-22.9)	5.0 (0.0-13.8)	13.1 (2.2-24.0)	11.0 (0.0-22.8)	478 2.1%
Other	0.8 (0.0-2.0)	24.6 (12.8-36.4)	0.7 (0.0-2.2)	15.0 (6.1-23.9)	7.9 (2.3-13.4)	1.5 (0.0-3.9)	22.0 (11.8-32.2)	27.5 (14.7-40.3)	1495 6.4%
Total	2.1 (1.0-3.3)	33.9 (28.0-39.7)	1.2 (0.3-2.0)	16.7 (12.1-21.4)	7.9 (6.4-9.4)	4.3 (2.7-5.9)	23.0 (20.1-26.0)	10.7 (7.9-13.6)	23150

$\chi^2_{42}=256$, p-value=0.0000

From Figure 9.9 it is clear that in males, head injuries are mainly caused by fights (46.9%) whereas in females domestic accidents (24.4%), fights (24.2%) and sport (37.8%) account for most head injuries. A high percentage of broken bones amongst males are caused by sport (20.1%) and approximately an even distribution exists between the other causes (10% to 15%), whereas in females a high percentage of broken bones are caused by domestic accidents (33.4%). It should be noted though, from Figure 9.9, there are overall less injuries amongst females (979 - broken bones, 523 - head injuries) than amongst males (1309 - broken bones, 1075 - head injuries) in the population.

FIGURE 9.9 GENDER SPECIFIC ANALYSIS BY CAUSE OF INJURY
(based on an 11 day recall period)



9.7 SUMMARY AND RECOMMENDATIONS

Despite the numerous results produced in this Chapter and the results based on child nutrition in Chapters 7 and 8, extensive research remains to be performed to identify aetiological factors in malnutrition, breastfeeding, disability, injury and maternal care. Notwithstanding this, the following broad relationships were identified in the various sections of this chapter.

Children with mothers that were uneducated had higher levels of malnutrition. Thus, for rural children, the beneficial effects of breastfeeding are outweighed by the negative effects of poverty and low levels of education of the mothers.

Factors such as lack of money, low level of education and residing in a rural area were all contributing factors to not attending ANC. A significantly higher percentage of previously pregnant urban mothers made use of a hospital/clinic/doctor for delivery than rural mothers. The main reason for not using a hospital/clinic/doctor for delivery in both the rural and urban areas was that the "birth came unexpectedly". The positive relationship between maternal care and urbanization is possibly due to the greater availability of appropriate facilities in urban than rural environments. However the effect of transport availability and costs in the rural areas on the attendance of ante-natal care, was not measured by the NH&NS. This might have a significant effect on attendance and it could be an important topic of future research.

For most disabilities the disability rate per 10000 population was significantly higher amongst males than females especially with regard to digit and limb amputations. These disabilities included amputation of both legs/feet, amputation of hand(s)/arm(s), severe deafness, amputation of fingers and toes, and fits. Working on mines was shown to be the main cause of these disabilities. Mining is second only to illness as a cause of disability in the whole population and amongst adult males it is the single biggest cause of disability.

The noticeable exception to this trend is the fact that females had almost double the rate of males in lameness/paralysis of both arms. Illness was the main cause of this disability and therefore the LMOH will need to further investigate the predominant types of illnesses that caused this disability. The disability rate was also significantly higher among uneducated than educated people, and significantly higher in rural areas than urban areas. The disability rate for all disabilities, except for "fits" which was fairly rare, increased significantly with age.

Ways of preventing the high prevalence of the following injuries also needs to be addressed: the high percentage of head injuries (39.4%), especially in males, caused by fights/assaults (46.9%); head injuries in females caused by domestic accidents (24.4%) and fights (24.2%); and the high percentage of burns (60%), especially in the 0-4 year age group (41%), mainly caused by domestic accidents. Domestic accidents and work/farm related injuries are presumably preventable with better education and training.

CHAPTER 10

SUMMARY

This thesis has highlighted the many difficulties experienced by the Lesotho National Household Health and Nutrition Survey. The fact that no results have been produced since the survey was conducted, is a clear indication of the difficulties experienced by the Lesotho Bureau of Statistics. Many observations were missing, others were erroneously deleted, and many errors appeared in other observations. An entire year of intensive data cleaning was necessary to remove the errors from the most important variables of the survey. An entire round of data (i.e. data collected from approximately 5000 households) proved to be unusable due to erroneous procedures that were employed during the data entering phase. Many of the errors could have been prevented or corrected had there been more efficient supervision and better management (see Chapter 3 for comments in this regard).

In Chapter 3 the various problem areas encountered in the NH&NS are identified, recommendations concerning better management are made, a data cleaning procedure designed for, and employed in the NH&NS is described, and the use of a palmtop computer during the interviewing phase is proposed. An example of a section from the NH&NS questionnaire is programmed on this palmtop. The capabilities and advantages of using this unit, as well as the source code to program the questionnaire are given. If this computerized interviewing process is adopted by a developing country, many data entry errors will be avoided, thus preventing a prolonged data cleaning phase such as that required in the NH&NS.

The simplified approach to the survey process is also presented in Chapter 3, where it is proposed that in developing countries only single round surveys should be conducted (Kroeger 1985). The thinking behind this is that a developing country often does not have the capability and expertise to conduct a survey with more than one round. The results in Chapter 3 suggest that no significant differences appear in the estimates over three rounds of data, for variables such as 'education level' that are not affected by seasonal patterns. However a seasonal effect is observed in malnutrition (and presumably other morbidity data). Therefore it would make sense to consider using only a single round survey for information that is not affected by seasonal changes. Before deciding to undertake a survey with more than one round factors such as the capacity of the survey team available; the need for such data; financial considerations; whether the population changes significantly in such short periods of time and time constraints should be considered.

Weighting the data to the population totals and compensating for missing responses can be a complicated and time consuming procedure. The weighting procedure of the NH&NS is described in Chapter 4. It might not always be necessary to compensate for item nonresponses. Recommendations are made in this regard based on results generated by using a resampling procedure. It was generally found that subgroup weighting for item nonresponse was preferable to imputation. These results can provide a developing country with a guideline to decide when it is necessary to compensate for item nonresponses. Two SAS programs to compensate for item nonresponses are included in the Appendices - the first using the technique of sample subgroup weighting, and the second using the technique of imputation.

The child nutrition section of the NH&NS is chosen for analysis throughout the thesis. In order to make a developing country collecting such data aware of the sources of errors and biases that can arise from using such data, approaches used in the literature and in the NH&NS to counter these problems are discussed in Chapter 5. Two data cleaning procedures specific to this type of data using interquartile ranges and a multivariate test based on the chi-square distribution, are introduced, in order to ensure more accurate results.

In Chapter 6 a program called STANDERR (written in SAS) for assisting the user in calculating estimates from weighted data from a complex survey design is described. Many well-known statistical packages exist (SAS and SPSS) that [illegible] not cater for a complex survey design such as the NH&NS. STANDERR therefore enables the user to use SAS and calculate standard errors, confidence intervals and design effect for data from a complex survey design. The process of portability is another approach that is considered for simplifying and speeding up the process of reporting the results of a survey. If applied correctly, the statistician analysing the data need not recalculate the sampling error for a variable, subclass, or survey round, once it has been calculated for another variable, subclass, or previous survey round. It was recommended that portability be applied only in the case of cross classes and across survey rounds in developing countries. Applying this procedure will therefore also simplify matters and save time.

In order to investigate the relationships between discrete variables from two-way or multiway tables of a complex survey design, two approaches are used - firstly, where the full knowledge of the covariance matrix under the sampling design is known (using the Wald statistic) and secondly where a correction factor is applied to the ordinary Pearson chi-square statistic. In the case of two-way or multiway tables, even though Fellegi's approach is considered conservative, this technique does not require knowledge of the full covariance matrix, and provided the design effects are known, other researchers will be able simply to recalculate the results for comparison. It is shown how not including the correction factor to compensate for a complex survey design alters the interpretation of the results. Therefore it **must** be included in the analysis of data from complex survey designs. For the analysis of two-way tables under simple random sampling, a test of homogeneity and independence are treated in the same way. Under a complex survey design though, these two tests require different adjustments and it is therefore important to distinguish between them. Examples of the two types of tests are also shown. Finally a log-linear model is fitted to a 2x2x2 contingency table to determine associations between age (below and above 24 months), breastfeeding (yes or no) and malnutrition (yes or no). It is shown that significant associations exist between all three factors. The associations indicate that children under 24 months benefit from breastfeeding, whereas over 24 months breastfeeding is really a sign of poverty which manifests itself through higher levels of malnutrition.

In Chapter 8 the Satterthwaite Adjusted Chi Square Test is used to determine whether a variable is to be included in the logistic regression model or not. This test accounts for clustering which, if not accounted for, the size of the standard errors of the coefficients are underestimated. In applying the logistic regression model to predict malnutrition among children it was shown that the education level of the mothers acted as an approximate 'surrogate' effect for place of residence in predicting malnutrition. Therefore, for rural children, the beneficial effects of breastfeeding are outweighed by the negative effects of poverty and low levels of education of the mothers. The process followed to find the best fitting model is discussed and the different ways to present this model are described.

Finally a chapter of results are presented in Chapter 9. These additional results should be of interest to the Lesotho Bureau of Statistics and Ministry of Health. In addition to the results on child nutrition produced in Chapters 7 and 8, this Chapter includes sections on maternal care, disabilities, and injuries. It is hoped that these results will be useful for future reference and assist in monitoring intervention programs, when similar health and nutrition surveys are again conducted in Lesotho.

For the future, improved procedures and better management needs to be implemented in a survey to be conducted by a developing country. The appropriate statistical techniques to analyze data from complex surveys such as the NH&NS must be applied or else misleading results will be produced.

Finally it is hoped that the recommendations made and techniques applied in this thesis will help simplify and improve various aspects of the survey process for a developing country to ensure that:

a) there is more accurate data;

b) the results are released sooner;

c) better management procedures are employed;

d) the appropriate techniques are used to weight data from a complex survey design;

e) the correct standard errors for a complex survey design are calculated ;

f) the appropriate techniques are used to analyze data from a complex survey design;

g) portability is used, where appropriate, to speed up the analysis and improve the quality of the results;

h) and survey statisticians will have a greater awareness of the inherent problems they might encounter in collecting data for large surveys with a complex survey design.

APPENDIX A

SAS PROGRAM CALLED STANDERR TO CALCULATE RATIO ESTIMATES, STANDARD ERRORS, CONFIDENCE INTERVALS AND DESIGN EFFECTS FOR RATIO ESTIMATES FROM A COMPLEX SURVEY DESIGN, DISCUSSED IN CHAPTER 6

```
* This program calculates the standard error, the design
* effect, and a 95% confidence interval, for a proportion from
* a multiway contingency table. The data is from a survey with
* a complex survey design. The input data must consist of the
* variables being estimated, the stratum number, the
* PSU number, a weighting factor which is optional, and the
* dimensions of the contingency table. For the example below,
* the analysis is performed on a two-way table of education
* level by sex.

data ggg; set allnew;
if sex='F' then ssex=1; if  sex='M' then ssex=2;
if educated^=.;

* The macro is executed according to the dimensions of the
* contingency table.

%macro serr;
%do i=1 %to 2;
%do j=1 %to 2;


data hhh;set ggg;
if educated=&i. and ssex=&j. then y=1;
else y=0;
if educated=2 or educated=1 or ssex=1 or ssex=2
then x=1;
else x=0;
if x=0 then y=0;
newt=ww4 ;
```

```
* Specification of the numerator and denominator of the
* proportion.

wx=newt*x;
wy=newt*y;

proc means sum noprint;
var wx wy x y educated ssex;
output out=tot sum=sumx sumy sum2x sum2y;

* Estimation of the proportion.

data jjj;set tot;
r=sumy/sumx;
r2=sum2y/sum2x;

* Estimation of the standard error taking the effect of
* stratification and clustering into account. See Chapter 5
* for more details.

data kkk; if _n_=1 then set jjj;
set hhh;
z=wy - r*wx;
proc sort;by strat psu;
proc means sum data=kkk noprint;
var z;
output out=lll sum=sumzp;
by strat psu;
proc means sum data=kkk noprint;
var z;
output out=nnn sum=sumzs;
by strat;
proc sort data = kkk;
by strat psu;

data mmm;
set kkk;
```

```
by strat psu;
if first.strat then count=0;
if first.psu then count + 1;
if last.strat and last.psu ;
proc sort data=lll;
by strat psu;
proc sort data=mmm;
by strat psu;
proc sort data=nnn;
by strat;
data ooo;
merge mmm lll;
by strat psu;

data ppp;
merge ooo nnn;
by strat;
proc sort;
by strat;

data qqq;
set ppp;
by strat;
if first.strat then sumzpsq=0;
sumzpsq+sumzp**2;
if last.strat;

data rrr;
set qqq;
total=(count*sumzpsq-sumzs**2)/(count-1);
proc means mean sum;
var total sumx r;
output out=sss mean=mntot mnsumx mnr sum=sumtot sumsumx sumr;
```

```
* The file 'Stderr.data' contains the variance-'ssqd', the
* standard error-'srssqd', the design effect-'deff', the
* confidence interval-'(lower;upper)', and the estimated
* proportion-'r'.

data ttt;
filename out 'Stderr.data';
file out mod;
merge sss (keep=sumtot mnsumx mnr)  jjj;
ssqd=sumtot/(mnsumx**2);
srssqd=sqrt(ssqd);
lower=mnr-2*srssqd;
upper=mnr+2*srssqd;
varsrs=(r2*(1-r2)/sum2x);
deff=ssqd/varsrs;
qq=symget('q');
put qq r deff ssqd lower upper;

%end;
%end;
%mend;

%serr;
run;
```

APPENDIX B

SAS PROGRAM CALLED ADDWEIGHT USED FOR SUBGROUP WEIGHTING TO COMPENSATE FOR ITEM NONRESPONSES DISCUSSED IN CHAPTER 4

```
* This SAS program is used to compensate for item
* nonresponses. The subgroup in this example is the PSU, and
* the missing responses are weighted for the variable
* "education".

libname dat '~';
data kkk;
set dat.subimp;

* The missing responses are identified. For stat=1 there is a
* valid response, and for stat=2 there is a missing
* observation.

if educate=1 or educate=2
then stat=1;
else stat=2;
proc freq ;
tables educate;

proc sort ;
by psu ;

proc freq;
tables stat/ out=frq;
by psu ;

* The number of missing observations in each PSU is counted,
* and a raising factor in each PSU - edr2=total2/total - is
* calculated

data dat.new2;
```

```
retain total 0;
set frq;
by psu;
if first.psu then total =count;
else total2=total + count;
if last.psu then do;
edr2=total2/total;
if stat=1 and percent>99.99 then edr2=1;
if stat=2 and percent>99 then stat2=2; else stat2=1;
output;
end;

* The new raising factor associated with each PSU is "kept",
* and merged to the original file containing the education
* information.

keep edr2  psu  ;
proc sort data=dat.subimp; by psu ;
proc sort data=dat.new2;by psu;

data dat.subimp;
merge dat.subimp(in=x)  dat.new2;
by psu ;
if x=1;
run;
```

APPENDIX C

SAS PROGRAM CALLED RANDALO TO CREATE 200 SAMPLES, ASSIGNING A VARIED PROPORTION MISSING IN EACH STRATUM, AS DISCUSSED IN CHAPTER 4

```
* This SAS program creates 200 samples, assigning a
* varied proportion missing in each stratum. In this
* particular example, 50% of the PSU's within each stratum are
* assigned 50% missing observations, and 50% of the PSU's are
* assigned 10% missing. For each sample created, the
* program STANDERR is called, to calculate an estimated
* proportion (educated persons over the age of 20 in this
* case), a standard error, confidence interval, and design
* effect.

* The macro is repeated for 200 samples.

%macro big;
%do q=1 %to 200;
data mmm;
set dat.subimp2;
if educated^=.;

* Reassigning the education variable.

if edattain='01' then edu='01' ;
else if edattain='02' then edu='02' ;
else if edattain='03' then edu='03' ;
else if edattain='04' then edu='04' ;
else if edattain='05' then edu='05' ;
else if edattain='06' then edu='06' ;
else if edattain='07' then edu='07' ;
else if edattain='08' then edu='08' ;
else if edattain='09' then edu='09' ; else edu='10';

if edu='01' or edu='02' or edu='03' then educated=1;
else if edu='04' or edu= 05' or edu='06' or edu='07' or
edu='08' or edu='09' then educated=2 ; else educated=.;

* The macro zz is performed on each of the PSU's.

%macro zz;
%do i = 1 %to 111;
data jj;
set mmm;
if newg=&i;
keep round psu hhno pn educated sex  fwt2 rurban;
proc means noprint n; output out=qq n=tot;
var educated;

data rr;set qq;
call symput('j',tot);
run;

* The following PSU's have 50% of their observations for the
* variable educated made missing.

proc sort data=jj;by round psu hhno pn;
data samp&i.;
if psu = '01101'
or psu = '01206'
or psu = '01302'
or psu = '02107'
or psu = '02115'
or psu = '02122'
or psu = '02202'
or psu = '02214'
or psu = '02307'
or psu = '03107'
or psu = '03114'
or psu = '03120'
or psu = '03207'
or psu = '03219'
or psu = '04101'
or psu = '04109'
or psu = '04116'
or psu = '04206'
or psu = '04218'
or psu = '04309'
or psu = '05107'
or psu = '05115'
or psu = '05121'
or psu = '05204'
or psu = '05216'
or psu = '06102'
or psu = '06109'
or psu = '06203'
or psu = '06305'
or psu = '06405'
or psu = '07301'
or psu = '07405'
or psu = '07417'
or psu = '08302'
or psu = '08413'
or psu = '09302'
or psu = '09317'
or psu = '10306'
or psu = '10320'
or psu = '10404'
or psu = '01502'
or psu = '02502'
or psu = '03501'
or psu = '03506'
or psu = '04501'
or psu = '04502'
or psu = '04504'
or psu = '04506'
or psu = '04510'
or psu = '04516'
or psu = '04522'
or psu = '05501'
or psu = '06501'
or psu = '07502'
or psu = '09501'  then do;
drop k ;
do k = 1 to &j./2 ;
iobs = int(ranuni(-1)*n) + 1;
set jj point=iobs nobs=n;
output;
end;
stop;
edatt2=.;
end;

* The remaining PSU's have 10% of their observations made
* missing.

else do;
drop k ;
do k = 1 to &j./10;
iobs = int(ranuni(-1)*n) + 1;
set jj point=iobs nobs=n;
output;
end;
stop;
edatt2=.;
end;

* Once the random assignment is completed, the ith bootstrap
* sample is saved.

keep round psu hhno pn edatt2 rurban;
proc sort ; by round psu hhno pn;
```

```
data nnn&i.;
merge jj samp&i. (in=x);
if x=1 then educated=edatt2;
drop edatt2;
by round psu hhno pn;
%end;

data all;
set %do k=1 %to 111;
nnn&k. %end;
%mend;
%zz; proc sort;
by round psu hhno pn;
proc sort data=dat.subimp2;
by round psu hhno pn;

data allnew;
merge dat.subimp2 all(in=x);
by round psu hhno pn;
if x=1;
run;

* The program STANDERR is called to calculate estimates,
* standard errors etc. on the ith bootstrap sample.

%include standerr;
%end;


* The statements signifying the end of the macro.

%mend;
%big;
```

APPENDIX D

SAS PROGRAM CALLED IMPUTER FOR IMPUTING MISSING OBSERVATIONS WITHIN A CLUSTER USING A RANDOM ASSIGNMENT PROCEDURE AS DISCUSSED IN CHAPTER 4

```
* This SAS program establishes which observations are missing
* in each PSU (the PSU is the subgroup in this case).
* Thereafter the missing observations are replaced by a random
* assignment of observations from the same PSU.


data mmm; set dat.imputer;
fwt2=round(fwt,1);

* The subgroups, which are the PSU's in this case, are
* identified for the variable 'newg'.

if psu = '01101' then newg = 1;
if psu = '01108' then newg = 2;
if psu = '01206' then newg = 3;
if psu = '01221' then newg = 4;
if psu = '01302' then newg = 5;
if psu = '01304' then newg = 6;
if psu = '02107' then newg = 7;
if psu = '02115' then newg = 8;
if psu = '02122' then newg = 9;
if psu = '02129' then newg = 10;
if psu = '02137' then newg = 11;
if psu = '02144' then newg = 12;
if psu = '02151' then newg = 13;
if psu = '02202' then newg = 14;
if psu = '02214' then newg = 15;
if psu = '02226' then newg = 16;
if psu = '02239' then newg = 17;
if psu = '02307' then newg = 18;
if psu = '02314' then newg = 19;
if psu = '03107' then newg = 20;
if psu = '03114' then newg = 21;
if psu = '03120' then newg = 22;
if psu = '03127' then newg = 23;
if psu = '03134' then newg = 24;
if psu = '03140' then newg = 25;
if psu = '03207' then newg = 26;
if psu = '03219' then newg = 27;
if psu = '03230' then newg = 28;
if psu = '04101' then newg = 29;
if psu = '04109' then newg = 30;
if psu = '04116' then newg = 31;
if psu = '04122' then newg = 32;
if psu = '04130' then newg = 33;
if psu = '04137' then newg = 34;
if psu = '04206' then newg = 35;
if psu = '04218' then newg = 36;
if psu = '04230' then newg = 37;
if psu = '04243' then newg = 38;
if psu = '04309' then newg = 39;
if psu = '04321' then newg = 40;
if psu = '05107' then newg = 41;
if psu = '05115' then newg = 42;
if psu = '05121' then newg = 43;
if psu = '05128' then newg = 44;
if psu = '05134' then newg = 45;
if psu = '05141' then newg = 46;
if psu = '05204' then newg = 47;
if psu = '05216' then newg = 48;
if psu = '05227' then newg = 49;
if psu = '05237' then newg = 50;
if psu = '06102' then newg = 51;
if psu = '06109' then newg = 52;
if psu = '06116' then newg = 53;
if psu = '06123' then newg = 54;
if psu = '06203' then newg = 55;
if psu = '06212' then newg = 56;
if psu = '06305' then newg = 57;
if psu = '06314' then newg = 58;
if psu = '06405' then newg = 59;
if psu = '06412' then newg = 60;
if psu = '07301' then newg = 61;
if psu = '07315' then newg = 62;
if psu = '07405' then newg = 63;
if psu = '07417' then newg = 64;
if psu = '07427' then newg = 65;
if psu = '07436' then newg = 66;
if psu = '08302' then newg = 67;
if psu = '08304' then newg = 68;
if psu = '08413' then newg = 69;
if psu = '08428' then newg = 70;
if psu = '09302' then newg = 71;
if psu = '09317' then newg = 72;
if psu = '09330' then newg = 73;
if psu = '09343' then newg = 74;
if psu = '10306' then newg = 75;
if psu = '10320' then newg = 76;
if psu = '10334' then newg = 77;
if psu = '10348' then newg = 78;
if psu = '10404' then newg = 79;
if psu = '10414' then newg = 80;
if psu = '01502' then newg = 81;
if psu = '02502' then newg = 82;
if psu = '02503' then newg = 83;
if psu = '02508' then newg = 84;
if psu = '02511' then newg = 85;
if psu = '03501' then newg = 86;
if psu = '03506' then newg = 87;
if psu = '03511' then newg = 88;
if psu = '03515' then newg = 89;
if psu = '04501' then newg = 90;
if psu = '04502' then newg = 91;
if psu = '04504' then newg = 92;
if psu = '04506’ then newg = 93;
if psu = '04510’ then newg = 94;
if psu = '04516’ then newg = 95;
if psu = '04522’ then newg = 96;
if psu = '04526’ then newg = 97;
if psu = '04532’ then newg = 98;
if psu = '04538’ then newg = 99;
if psu = '04542’ then newg = 100;
if psu = '04548’ then newg = 101;
if psu = '04553’ then newg = 102,
if psu = '04559’ then newg = 103;
if psu = '05501’ then newg = 104;
if psu = '05505’ then newg = 105;
if psu = '06501’ then newg = 106;
if psu = '06505’ then newg = 107;
if psu = '07502’ then newg = 108;
if psu = '08502’ then newg = 109;
if psu = '09501’ then newg = 110;
if psu = '10501’ then newg = 111;


if educated=1 then educa=1;
if educated=2 then educa=2;
if educated=. then educa=.;

proc freq;
tables educa;

* The macro is performed for each of the 111 PSU’s.
* The missing responses in each PSU are established.

%macro zz;
%do i = 1 %to 111;

data jj;
set mmm;
```

```
if newg=&i and educa=. ;
keep round psu hhno pn educa sex fwt2;
proc means nmiss;
output out=qq nmiss=tot;
var educa;

data rr;set qq;
call symput('j',tot);
run;
proc sort data=jj;
by round psu hhno pn;


* The valid responses in each PSU are established
* and assigned to the variable edatt2

data kk;
set mmm;
if newg=&i and educa ^=.;
edatt2=educa;
keep round psu hhno pn educa edatt2 sex fwt2;

* The missing observations in each PSU are randomly assigned a
* valid response from the same PSU

data sample;
drop i ;
do i = 1 to &j.;
iobs = int(ranuni(0)*n) +1;
set kk point=iobs nobs=n;
output;
end;
stop;
keep round psu hhno pn edatt2;

proc sort ;
by round psu hhno pn;
```

```
data nnn;
merge jj sample;
if educa   =. then educa=edatt2;
drop edatt2;

data ppp&i;
set kk nnn;
proc freq;
tables educa;
%end;

* The newly created data set is merged with the original
* one.

data all;
set %do k=1 %to 111;ppp&k. %end; ;
%mend;
%zz;
proc sort;
by round psu hhno pn;

proc freq;
tables educa;

proc sort data=dat.si4;
by round psu hhno pn;

data dat.si5;merge dat.si4 all;
        by round psu hhno pn;
```

APPENDIX E. THE LESOTHO NATIONAL HOUSEHOLD HEALTH AND NUTRITION SURVEY

CONFIDENTIAL

STAT/NHNS/1 BUREAU OF STATISTICS

NATIONAL HOUSEHOLD HEALTH AND NUTRITION SURVEY

QUESTIONNAIRE

ROUND 4

Form A - Household Questionnaire
Form B - Age Determination
Form C - Morbidity Questionnaire
Form D - Maternal Care Questionnaire
Form E - Nutrition Questionnaire
Form F - Health Education Questionnaire

(FILL IN THIS PART FIRST)

Rural (1) Urban (2) □ PSU □□□□□□

Sample Household No. □□

Village ....

Name of Head of Household ....

Village Chief ....

Callbacks

	Day		Mon.		Year	
1st						
2nd						
3rd						
Interview's Date						

No. of Books Used ....

Enumerator's Name .... Code .....

Supervisor's Name ....

FORM A

ASK ALL HOUSEHOLD MEMBERS

Line No.	Name	Relationship to head of household	Present or absent	Sex	Age		Marital status	Education	Main (usual) Occupation	Ever worked on mines	Additional forms to be used			
	Start with head of household even if absent. PROMPT: Is there anyone who usually lives here but is presently in hospital? PROMPT: Were any children born in the last year who have already died? Also include visitors staying for the major part of last 2 weeks.	01 Head 02 Spouse 03 Son/daughter of H/H 04 Spouse of son/daught. 05 Gr-grand/Grand child 06 Parent of head/spouse 07 Brother/sister of head/spouse 08 Nephew/neice of head/spouse 09 Other relative 10 Domestic employee 11 Not related 12 Other	0 Dead now 1 Present for most of last 2 weeks and present now, or returning in 7 days. 2 Absent now - present for most of last 2 weeks, will not return in 7 days. 3 Absent in Hospital or trad.healer 4 Absent for most of last 2 weeks 5 Not part of H/H	Code: M = male F= Female	6 yrs and over -------- Completed years	Under 6 years -------- Age in completed MONTHS (from form B)	M Married/ living together N Never married D Divorced/ seperated W Widowed	Give highest standard passed ---------------- 00 Not applicable (below 6 yrs) 01 No formal education 02 Standard 1-4 03 Standard 5-7 04 JC/LPTC/equ 05 COSC/equiv 06 Post COSC 07 Technical/ vocational 08 Degree or higher 09 Other	00 Not applicable (below 10 yrs) 01 Employer 02 Own account worker/farmer 03 Unpaid family worker 04 Regular wage/ salary earner 05 Casual worker 06 Unemployed 07 Student 08 Homemaker 09 Pensioner, retired 10 Other	To be asked of males over 15 years. -------------- 0 Not applic. 1 Never worked on mines 2 Worked on mines but never worked underground 3 Less than 5 years underground 4 5 to 10 yrs underground 5 More than 10 years underground	Form B (Age) --------- Tick (✓) if child under 6 yrs	Form C (Morbi-dity) --------- Tick (✓) if Col 4 has a 1,2 or 3	Form D (Maternal Care) --------- Tick (✓) if Female age 15-49 AND Col 4 = 1	Form E (Nutrition) --------- Tick (✓) if child under 60 months (Col 7) AND Col 4= 1
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1														
2														
3														
4														
5														
6														
7														
8														
9														
10														

Main source of household income

16. What is the main source of income of this household?
(*Circle one answer)

1 Subsistence farming
2 Cash-Cropping or Sales of Livestock
3 Business income
4 Cash remittances from migrant workers
5 Wages or salaries in cash (not migrant workers)
6 Other (specify)..

17a. Does this household own any fields? Yes No

b. If yes: How many fields? ______________

18a. Does this household own any cattle, sheep or goats? Yes No

b. If Yes: How many CATTLE does it own? ______________

Housing

19. How many rooms (counting all buildings) does the household have? ________

Health facilities

Nearest known facility or facility usually used	a. Name or Place If none write NA If don't know, write DK	b. How long to walk 1 way (If "don't know", enter 99 (-?))		c. How much does transport cost (for round trip)		d. *SUPERVISORS Code distance in kilometers
		Hours	Minutes	Maluti	Lisente	(000 - 999)
20. Hospital (Name)						
21. Health centre (Place)						
22. Mobile clinic/outstation (Place)						
23. Traditional healer (Place)						
24. Private doctor (Name)	______________					
(Place)	______________					

25. Is there a Village Health Worker for your village/area?

No Don't Know Yes

26. What is her/his name?
(* if don't know, write DK) ______________

27. How many times has he/she visited your home in the last month? __________

INTERVIEWER'S CHECK
28. Is there a trained VHW for this area? (Y, N) ____

29. If yes, is VHW name in Q26 correct? (Y,N,DK,X) ____

NEXT FORM

Supervisor's check
1. Check private doctor's registration.......... ____

2. Calculate and code distances (questions 18 -22 d) ____

FORM B AGE DETERMINATION

To be completed for every child under the age of 6 years (72 months)

Child's Name	Line No. (from form A)	Is there a document with written date of birth? (Y/N) No ----> Use Mother's recall and local events calendar Yes ---->	Type of document? A. Growth chart B. Bukana C. Birth certificate D. Baptismal cert. E. Other (specify)	----> Date of birth			Age in months (from age calculator)
				Day	Month	Year	
1	2	3	4	5	6	7	8

FORM C

1 Name:____

2 * LINE NUMBER (from form A):__________

3 Respondent (CIRCLE ONE CODE): 1 Self 2 Mother 3 Grandmother
4 Spouse 5 Father 6 Sister
7 Maid 8 Other guardian, relative

4 * IN CALENDAR BELOW, MARK AN X IN THE ROW "TODAY" IN THE BOX INDICATING THE DAY TODAY.
* DELETE ALL BOXES IN ALL ROWS TO THE RIGHT OF X WITH A LARGE W.
* COUNT THE NUMBER OF DAYS FROM THE PREVIOUS SUNDAY TO TODAY INCLUSIVE:_______
(Note: answer should be between 9 and 15)

5 Ho tloha ka Sontaha se ka pele ho se fetileng, ke hore matsatsi a ____ a fetileng na u kile oa kula?

YES

Please describe the type of illness or injury. What were the main problems?

* AFTER SPONTANEOUS ANSWERS HAVE BEEN GIVEN →

NO

I would just like to check whether you have had any of the following problems:
* PROMPT ONLY THE UNCIRCLED HEADINGS

(*CIRCLE THE NUMBER FOR EACH SYMPTOM OR ILLNESS. ALSO CIRCLE THE HEADING OF THE GROUP WHERE THE SYMPTOM APPEARS.)

a. Khohlela
1. Khohlela
2. Sefuba
3. Lefuba
4. Letsoaha
5. Sejeso
6. Mokhokho-thoana
7. Letsoejana
8. Tematoana/'awetso

b. Letsoilo Mohlala
1. Mohlala
2. Letsoilo
3. L.lafubelu
4. L.letala
5. L.lesehla
6. L.metsi
7. L.ka mariana
8. Kokoana

c. Tsebe
1. Tsebe tse bohloko
2. tse bolalu
3. Mokhasa
4. Kokoana

d. Mahlo
1. Jeoa ke Mahlo
2. Mahlo a mafubelu
3. Mahlo a llang
4. Mahlo a melaka

e. Letlalo
1. Lekhopo
2. Maselese
3. Liso/Seso
4. Lethopa
5. Mosolo
6. Khopole
7. Litlapeli
8. Sahloba
9. Motsokapere/Ngoae

f. Mocheso
1. Moch..o
2. Feberu

g. Akheha Sethoathoa
1. Akheha
2. Sethoathoa

h. Leino
1. Jeoa ke leino
2. Leino le sekoti
3. Haranene a tsoa mali

i. Theoha Meleng
1. Otile
2. Theoha Meleng

j. General acute
1. Ngoana oa Sentsoeng
2. Phuana
3. Mokola
4. Kokoana
5. Thokolosi

k. Mathetha a Basali
1. Mathetha a matsatsi
2. Senyehetsoe
3. Mathatha a pelehi/pepa
4. Seso se sengata
5. Mokaola
6. Popelo
7. Letheka
8. O na le noka
9. Nyopa

l. Mathatha a Banna
1. Mokaola
2. Seso
3. Metsi a bohloko
4. Ha ke tsohaloa

m. Lehlaba
1. Hlohong
2. Thekeng
3. Soholoholo
4. Manonyeletso
5. Masapo
6. Mosifa/cramps
7. Mothapa oa Leoto
8. Pakeng tsa Mahetla

n. Mpeng
1. Mala
2. Liso ka mpeng
3. Mahlaba ka mpeng
4. Chachametsa
5. Pipitlelos
6. Manyoua
7. Lehlatso
8. Nyoko
9. Chefu
10. Lefetsoaneng

p. General - Chronic
1. Mali a fokola
2. Mele oa fokola
3. Mokhathala
4. High blood
5. Lefu la tsoekere
6. Lefu la pelo
7. Borurisi
8. Molikoalikoane

q. OTHERS:____

6 * IF THERE WERE NO ILLNESSES SINCE THE PREVIOUS SUNDAY, SKIP TO QUESTION 11.

7 * GROUP THE CATEGORIES INTO ILLNESSES AND LABEL THEM A, B AND C IN THE BOX NEXT TO THE APPROPRIATE HEADINGS.
- (If there were more than 3 illnesses, choose only the 3 most important illnesses. If all symptoms belong to the same illness, label them all A.)

* START WITH ILLNESS A: ASK QUESTIONS 8, 9 & 10. THEN ASK 8, 9, & 10 FOR ILLNESS B, AND FINALLY FOR C.

8 When did ...(illness A/B/C)... end? (* Mark X to show the last day in the row for this illness.)
(* If illness still present today, place X in the column for today.)

9 How long have you had ...(illness A/B/C)...?
(* If illness lasted more than 4-weeks, put X in correct column on left, and join Xs with a line.)
(* If less than 4 weeks, mark X for each day of illness going backwards from last day..

10 Did this illness begin *before, on, or after* the previous Sunday?
* COMPARE THIS ANSWER WITH THE CALENDAR AND CORRECT IF NECESSARY.

	More than 1 year	More than 3 mnths up to 1 year	More than 4 weeks up to 3 mnths	Up to 4 weeks								Previous Sunday														
					Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su
TODAY	//////	/////////	/////////	///////	//	//	//	//	//	//	//	//	//	//	//	//	//	//	//							
ILLNESS A																										
ILLNESS B																										
ILLNESS C																										

11 *IF INJURIES HAVE ALREADY BEEN MENTIONED AND ALL RELATED QUESTIONS ASKED, SKIP TO QUESTION 15.

12 Na u kile oa tsoa kotsi e sale e le ho *tloha Sontaheng se kapele ho se fetileng?* (* IF NECESSARY, PROMPT types of injury)

No

Yes

INJURIES D
(* Circle codes: max = 3)

a. Broken bone
b. Torn skin or flesh
c. Burns
d. Head injury
e. Sprain or bruise
f. Eye injury
g. Other

13 How was the injury caused?
(* Circle 1 code only)

1. Traffic accident
2. Domestic accident
3. Mine accident
4. Other Work/Farming accident
5. Fight/assault
6. Horse
7. Playing, sport
8. Other (specify)__________

14 *CHECK: Did this happen since the previous Sunday? (* circle answer)

Yes No

15 *CHECK: ARE ANY ILLNESSES MARKED IN THE CALENDAR? YES -----> SKIP TO QUESTION 16.

NO

* WERE THERE ANY INJURIES (Question 12)? YES -----> SKIP TO QUESTION 16

NO ------> SKIP TO QUESTION 27

16 Lefu kapa kotsi e neng e u totetse haholo ho feta tse ling ke efe? *PROMPT illnesses A, B, C & D.
(*A, B, & C represent the illnesses as you have labelled them in question 7. D represents injuries)

A B C D

17 Concerning this ..(illness/injury).. for how many days did you have to spend most of the day in bed? (*Write no. of days. If less than 1 day, put 0.) ______

18 Concerning this ..(illness/injury).. for how many days were you unable to do your normal activities (e.g. housework, school, farming) ?
(*Write no. of days. If less than 1 day, put 0.) ______

19 What did you do about this illness/injury and where did you go for treatment?

IF NECESSARY, PROMPT: Sometimes when I get ill I go to Ngaka ea Sesotho or Mopostola. Did you go to a Ngaka ea Sesotho or Mopostola for this illness?

(* Circle the number for each source of care. Max = 5)

1. Nothing
2. Self care with sesotho medicines
3. Self care with medicines from shop/chemist/pharmacy
4. Ngaka ea Sesotho ....
5. Mopostola/Spiritual healer..........................
6. Trained village health worker
or trained traditional birth attendant...............
7. Private doctor (name________________________)
8. Private Nurse ....
9. Mobile clinic, clinic out-station ....
10. Clinic, health centre, dispensary..................
11. Hospital ....
12. Other (specify)....................................

20 Maikutlo a hau ka afe mabapi le litsebeletso ka kakaretso? (*DO NOT PROMPT)

(* Code: Write letter in line of facility to which it refers. Max of 3 codes per facility may be entered.)

a Fees too high	f Treatment did not help
b Runs out of drugs	g Criticises competence of staff
c No food scheme	h Criticises attitude of staff
d No injection given	i Satisfied, getting better
e Long wait before seen	j Other (specify)

21 *IF PRIVATE DOCTOR (7), NURSE (8), CLINICS (9,10) OR HOSPITAL (11) WERE MENTIONED SKIP TO QUESTION 23.

22 There are many reasons why people do not go to a health facility for every problem. Why did you not go to the clinic or hospital for this ....(illness/injury)... ?
(* Circle number. Max = 3)

1. Illness not serious enough
2. Traditional medicines and treatment from Ngaka ka sesotho more appropriate for this illness
3. Believes bewitched or possessed - Ntho tsa basotho
4. Believes treatment from clinic does not help
5. Distance to clinic/hospital too far
6. Bad weather prevented going to clinic
7. Clinic does not always have drugs
8. Have to wait too long to be seen
9. Attitudes of staff
10. Husband or mother or others discouraged/prevented me from going
- Employer would not allow me time to go
11. No Money ---------->* PROMPT: What is it that you do not have enough money for?
12. Transport
13. Fees
14. Other (specify)______________
15. Other (specify)______________________________

23 * CHECK BACK TO QUESTION 5 : DID .... REPORT HAVING A COUGH?

No Yes

24. * CHECK FROM CALENDAR: DID THE COUGH LAST MORE THAN 4 WEEKS?

No Yes

25. I would like to check something about the cough you have had.
Did you EVER go to a clinic or hospital about this cough?

No Yes Don't know

26. Ha u ne u ile cliniking kapa . .petlele u ile oa koptjoa ho fana ka sekhohlela?

No Yes Don't Know

27 Have you slept in hospital as a patient (for at least 1 night) in the last year i.e. since last ..(name of month)..?

No Yes Don't Know

28 *IF DISABILITIES HAVE ALREADY BEEN MENTIONED AND ALL RELATED QUESTIONS ASKED, GO TO NEXT FORM.

29 Na ...Lebitso.. u na le kholofalo ea mofuta ofe kapa ofe? ([illegible], bomumu, botholo, kapa matsoho kapa menoana e khaohiling, kapa bokuli ba kelello)

No Yes

30 Sehole/kholofalo
(* Circle letter. Max = 3)

a. Amputation of 1 or more fingers	i. Lame / paralysed - both legs
b. Amputation of hand(s)/arm(s)	j. Blind (total)
c. Amputation of 1 or more toes	k. Blind (partial)/short sighted
d. Amputation of 1 foot/leg	l. Severe deafness
e. Amputation of both feet/legs	m. Speech problems
f. Lame / paralysed - 1 arm	n. Mental illness, retardation
g. Lame / paralysed - both arms	o. Fits - Sethoathoa (for many years)
h. Lame/paralysed - 1 leg	p. Other (specify)________

31 For how long has ..(name).. been disabled?
(* Circle 1 no. only)

1. Since birth
2. Since childhood
3. Since adulthood
4. Don't know

32 What was the cause of the disability?
(* Circle 1 number only)

1. Born disabled	7. Fight/assault
2. Illness	8. Playing, sport
3. Traffic accident	9. Horse
4. Domestic accident	10. Witchcraft/Moea oa Bothuoela
5. Mine accident	11. Unknown
6. Other work/farming accident	12. Other (specify)...................

* NEXT FORM

FORM D MATERNAL CARE QUESTIONNAIRE

* TO BE ASKED OF ALL WOMEN BETWEEN 15 AND 49 YEARS OF AGE

1. Name ______________________________

2. * LINE NUMBER (from form A) _______

3. Na u lebeletse ngoa? (* circle one answer)

No Don't know Yes

4. U nako e kae u le mokhachane?
(* Write number of months
or DK if don't know) __________

*SKIP TO QUESTION 8

5. Na u kile oa ba le ngoana ea hlahileng a phela likhoeli tse leshome le metso e 'meli tse fetileng? Ka mantsoe a mang kamor'a hlakubele selemong sa 1988. (*Circle one answer)

Yes No

6. Ho ikhotsofatsa: 'm'e u re hohang ha u so be le ngoana ea ileng a hlaha a ba a feta?

Yes No

*GO TO NEXT FORM.

7. What was the date of birth of the most recent child (even if that child died)?

Day_______ Month__________ Year_________

8a. Have you had a previous live birth?

No Yes

8b. Is this previous child now alive?

Yes No

8c. How many live births have you had altogether?
(*Write number)

____

* GO TO NEXT QUESTION 9

9. During the most recent pregnancy, (* i.e. the present pregnancy if pregnant, the most recent live birth if not pregnant), did you receive any special care for the pregnancy (ante-natal care or ANC)?

No → 11   Yes → 10   ☐

10. Where did you receive this care? (* circle one no.only)   ☐

1 Clinic (health centre, dispensary) or hospital → 12
2 Private doctor, private nurse.
3 Traditional (Mopostola, ngaka ea Sesotho, Mopepisi → 11
4 Trained VHW, trained traditional birth attendant i.e. Mopepisi who has been trained at the clinic → 11
5 Self care, family → 11
6 Other (specify).... → 11

11. There are many reasons why people do not go to a clinic during pregnancy. Why did you not go? (*Maximum 2 answers)   ☐☐

a. Clinic too far
b. No money to go to clinic
c. Sickness needs ngaka ea sesotho
d. Other (specify):

12. How many times did you go for ANC?

________   ☐☐

13. IF PREGNANT ——> Where do you intend to have the baby delivered?
IF NOT PREGNANT ——> Where was the child delivered?
(* circle one no. only)   ☐

1 Hospital → 16
2 Clinic (health centre, dispensary) → 16
3 Private doctor's/nurse's surgery → 16
4 Home → 14
5 Other (specify)...................... → 14

14. Who assisted (will assist) in the delivery? (*circle one no.)   ☐

1 Mopepisi who has trained at the clinic (Trained TBA) or VHW
2 Relative
3 Friend
4 Other (specify)..................................

15. There are many reasons why women do not deliver in a health facility? Why did/will you not? (*Circle maximum 3 answers)   ☐

a. Lack of money   b. Facilities too far
c. The birth came unexpectedly   d. Clinic was closed
e. Prefer to deliver at home
f. Do not find it necessary to deliver at clinic
g. Attitude of clinic/hospital staff
h. Tradition is to deliver at home to learn how to deliver others
i. Delivered at home so that could bury placenta
j. Other (specify):

16. Did/will you and the child visit any clinic or hospital during the 2 months after the child's birth?

Yes → NEXT FORM   No → 17   Don't Know → 17   ☐

17. There are many reasons why women do not go to the clinic during the period after the baby is born?
Why did/will you not visit the clinic? (*Max. 3 answers)   ☐

a. Did not feel sick/ mother & baby were fine
b. No money
c. Child passed away
d. Child had already received first immunisation
e. Too young to visit clinic
f. Mother was ill
g. No reason
h. Other (specify):

* NEXT FORM

FORM E NUTRITION QUESTIONNAIRE

This form must be asked of all children less than 60 months old.
Ask mother if possible! If mother not available ask major carer.

1. Child's name:______________________________

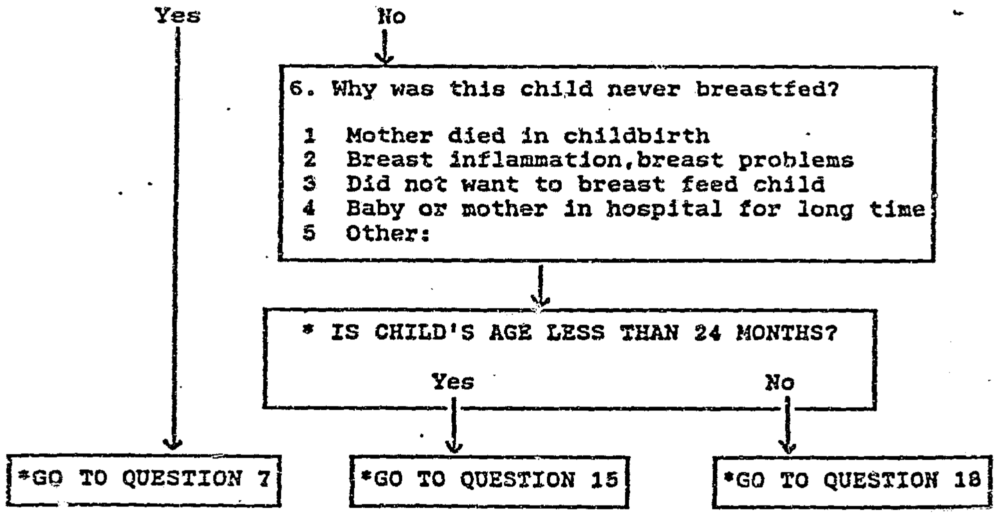
2. * LINE NO. FROM FORM A:__________

3. * AGE IN COMPLETED MONTHS FROM FORM A: __________

4. Maobane ke bo mang ba neng ba fepa ngoana?
(*Tsoaea ba sa feteng bobeli)

a Mother	b Grandmother	c Aunt	d Maid
e Father	f Sister	g OTHER:.............	

5. Was this child ever breastfed?



Yes

No

6. Why was this child never breastfed?

1 Mother died in childbirth
2 Breast inflammation,breast problems
3 Did not want to breast feed child
4 Baby or mother in hospital for long time
5 Other:

* IS CHILD'S AGE LESS THAN 24 MONTHS?

Yes

No

*GO TO QUESTION 7

*GO TO QUESTION 15

*GO TO QUESTION 18

C4.

7. Is this child still breast feeding?

Yes

No

8. How old was the child when he sucked from the breast for the last time?

(*Age in completed months)________

9. Why did you stop feeding the child at that age? (*Circle one letter).

a. Child old enough
b. Child refused to suck
c. There was no milk in the breasts
d. Father wanted to have sex
e. Parents wanted another child
f. Became pregnant
g. Mother was sick
h. Mother was working
i. Mother was going to school
j. Mother was tired of breastfeeding
k. Mother was trying to run away from the father
l. Child taken from mother because father did not like it
m. Don't know
n. Other:..

10. *IS THE CHILD'S AGE LESS THAN 24 MONTHS?

Yes

No

*GO TO QUESTION 18

11. What was the first thing fed to this child after birth?

a. Breast milk - PROBE OTHERS TO BE SURE
b. Plain water
c. Sugar water
d. Infant formula
e. Other (specify).................................
f. Don't know

12. How soon after birth did this child receive breast milk for the first time?

a. Immediately (< 1 hour)   b. Within a few hours (1-6 hrs)
c. Within one day (7-24 hrs)   d. On second day (25-48 hrs)
e. On third day or after   f. Don't know

13. *IF CHILD IS NO LONGER BREASTFEEDING (CHECK FROM QUESTION 7) SKIP TO QUESTION 15.

14. Is this child EXCLUSIVELY breastfed? (*Circle Yes only if the child has received nothing but breast milk since yesterday morning, not even water.)

No

*GO TO QUESTION 15

Yes

*GO TO QUESTION 18

15. Now I would like you to tell me about all the foods and drinks your child has had from the time he/she woke up yesterday morning until before he/she woke up today.
(*Tick off the respondent's replies in the table keeping all the food eaten in one meal in the same column, and a different column for each meal.)

Foods		Meal or snack number							
		1	2	3	4	5	6	7	
a1. Metsi	PROMPT								a1 └─┘
a2. Metsi a tsoekera	PROMPT								a2 └─┘
a3. Tee	PROMPT								a3 └─┘
a4. Lero litholoana	PROMPT								a4 └─┘
a5. 'Oros', squash, coke etc.	PROMPT								a5 └─┘
a6. Lebese - la khomo	PROMPT								a6 └─┘
a7. Lebese - la phofo	PROMPT								a7 └─┘
b1. Papa - Poone									b1 └─┘
b2. Papa - Mabele									b2 └─┘
b3. Lesheleshele - Poone									b3 └─┘
b4. Lesheleshele - Mabele									b4 └─┘
b5. Motoho o Ritetsoang - Poone	PROMPT								b5 └─┘
b6. Motoho o Ritetsoang - Mabele	PROMPT								b6 └─┘
b7. Motoho/ Mahleu (Other porridges)									b7 └─┘
c1. Bohobe									c1 └─┘
c2. Biskiti									c2 └─┘
d1. Moroho (greens)									d1 └─┘
d2. Lierekisi, linaoa									d2 └─┘
d3. Mokopu									d3 └─┘
d4. Meroho e meng (tomato, litapole lohoate etc.)									d4 └─┘
d5. Litholoana									d5 └─┘
e1. Nama									e1 └─┘
e2. Hlapi									e2 └─┘
e3. Mahe									e3 └─┘
f1. Commercial baby food - Cereal (Nestum)									f1 └─┘
f2. - 'Purity'									f2 └─┘
f3. 'SIMBA', Lipompong	PROMPT								f3 └─┘
f4. Other foods:........................									f4 └─┘
f5. Other foods:........................									f5 └─┘

likoete

16. *COUNT NUMBER OF MEALS AND SNACKS I.E. NUMBER OF COLUMNS WITH TICKS. THEN CHECK WITH THE RESPONDENT THAT THIS IS THE CORRECT NUMBER OF SEPARATE MEALS AND SNACKS FED TO THE CHILD YESTERDAY.

NUMBER OF MEALS AND SNACKS:__________ └─┘

17. Do you own a baby bottle (botlolo e fepang ngoana)?

Yes No Don't know └─┘

18. Does this child receive any food donations?

Yes No Don't know └─┘

19. Arm circumference: └─┴─┘,└─┘ cms

(Arm circumference was not taken because:..........................
....)

20. Weight: └─┴─┘,└─┘ kg

(Weight was not taken because:.....................................
....GO TO Q24)

21. Which of the following was the child suspended in during the weighing?

a. Plastic trousers b. Cloth trousers └─┘
c. Cloth sling d. Other (specify).....................

22. Clothes worn during weighing: (*Circle the codes of all that apply)

1. None 2. Pre-weighed blanket └─┴─┴─┘
4. Underpants 8. Undershirt
16. Petticoat 32. Light diaper/nappy
64. Heavy diaper/nappy 128. Other (specify)..................

23. Was the weight at the usual standard of accuracy?

Yes No └─┘

Explain:...

....

24. Length: └─┴─┴─┘,└─┘ cms

25. Was the length at the usual standard of accuracy?

Yes No └─┘

26. Explain:...

....

27. Edema of the feet: (*More than one answer may be circled)

None Left foot Right foot Unsure └─┘

28. Persons who have answered most of these questions:
(*More than code may be circled)

1. Mother 2. Grandmother 4. Aunt 8. Maid └─┴─┴─┘

16. Father 32. Sister 64. OTHER:..............

FORM F. AIDS AND HEALTH EDUCATION QUESTIONNAIRE

1. Na u se u kile ua utloela ka lefu la AIDS?
(*Circle one answer)

E Che ————————> *Go to question 9

2. Motho a ka tsoaroa ke lefu lee nako e kae pele a ka bontsa matsoao a ho kula? (*Answer number of months or years, or DK if don't know.)

____

3. Ke mokhoa ofe o tsebahalang oo lefu la AIDS le ka tsoaetsanoang ka oona? (*Place one tick (✓) in first column of table. Do not prompt except to clarify answers.)

	Most common (One ✓ only)	Others (Tick all
a. Ho arolelana likobo la motho ea nang le kokoana-hloko ea AIDS............		
b. Ho arolelana likobo le motho ea tsoaroeng ke lefu la AIDS................		
c. Ho arolelana likobo le batho ba bangata..................................		
d. Ho arolelana likobo le melata...		
e. Ho arolelana likobo le batho ba tsoang/hlahang RSA		
f. Ho arolelana likobo le batho ba boitsoaro bo boba (prostitutes/matekatse)..		
g. Ho Jana maotoana le motho ea tsoaroeng ke lefu la AIDS....................		
h. Ho jana maotoana le motho ofe kapa ofe..................................		
i. Ho fuoa mali a nang mang la kokoana-hloko ea AIDS..........................		
j. Ho sebelisa onts o sebelisitsoang ka motho ea tsoaroeng ke lefu la AIDS. kapa ea nang le kokoana-hloko ea AIDS............................		
k. Ho phunya litsebe ka nale e sebelisitsoeng ke motho ea nang le le kokoana-hloko ea AIDS..........................		
l. Ho sebelisa nale naie e sebolisitsoeng ke motho ea tsoaroeng ke lafu la AIDS...		
m. Ho sebelisa lehare le sebelisitsoeng ke motho ea nang le kokoana-hloko ea lefu la AIDS kapa ea tsoareng ke lefu la AIDS.....................		
n. Ho phatsa ka lehare la sebelisitsoeng ka motho ea tsoareng ke lefu la AIDS kapa ea nang le kokoana-hloko ea AIDS......................		
o. Ho alimana liphahlo le motho ea tsoaroeng ke lefu la AIDS................		
p. Ho alimana/arolelana lipitsa/lijana le motho ea tsoaroeng ke lefu la AIDS..		
q. Ho sebelisa ntloana le motho ea tsoeroang ke lefu la AIDS................		
r. Ho kopanela borashe ba mano le motho ea tsoereng ke lefu la AIDS..........		
s. Ho aka (kissing) motho ea tsoaroeng ke lefu la AIDS......................		
t. OTHER (specify)..		
u. OTHER (specify)..		

4. Ke mekhoa efe hape eo AIDS e ka tsoaetsanang ka eona? (*Place a tick (✓) in the 2nd column in the above table next to every method mentioned. Do not prompt except to clarify answers.)

5. Ke mokhoa ofe o [illegible]bahalang oo motho a ka o sebelisang ho qoba lefu la AIDS?
(*Do not prompt except to clarify answers.
Place only one tick in first col. of table below.)

	Most common (One ✓ only)	Others (Tick all)
a. Ho ba le motsoalle a le mong		
b. Ho qoba karolelano ea likobo le:		
1) motho ea nang le kokoana-hloko ea AIDS		
2) motho ea tsoeroang ke AIDS		
c. Ho qoba ho arolelana likobo le batho ba boitsoaro bo bobe (prostitutes/matekatse)		
d. Ho qoba ho ja maotoana le:		
1) motho mang kapa mang		
2) motho ea.tsoeroeng ke AIDS kapa ea nang le kokoana-hloko ea AIDS		
e. Ho sebelisa khohlopo		
f. Ho sebelisa mekhoa e meng ea thibelo ea pelehi		
g. Ho qoba ho aka motho ea tsoeroeng ke AIDS kapa motho ea nang le kokoana-hloko ea AIDS		
h. Ho se sebelise ente e sa hloekisoang		
i. Ho se sebelise nale (pin) e sebelisitsoeng ke motho ea tsoeroeng ke AIDS kapa ea nang le kokoana-hloko ea AIDS.		
j. Ho se phatse ka lehare le sebelisitsoeng ke motho ea tsoeroeng ke AIDS kapa ea nang le kokoana-hloko ea AIDS		
k. Ho qoba ho beola ka lehare le sebelisitsoeng ke ea nang lo kokoana-hloko ea AIDS		
l. Ho se aparelane liphahlo le motho ea tsoeroeng ke AIDS		
m. Ho se kopanele lijana le motho ea tsoeroeng ke lefu la AIDS kapa ea nang le kokoana-hloko ea AIDS		
n. Ho se kopanele borashe bo hlatsoang meno		
o. Ho tsamaea tlhatlhobo ea bophelo		
p. Other (specify)		
q. Other (specify)		

6. Ke mekhoa efe hape eo motho a ka e sebelisang ho qoba lefu la AIDS? (*Do not prompt except to clarify. Place a tick for each answer in the second column of the table above.)

7. Taba tsee tsa ho qoba AIDS u li utloile kae?
(*Circle the number for each answer given)

1	Radio Lesotho	32	Family planning worker
2	Radio Sesotho	64	Friend, relative
4	Doctor	128	Newspaper
8	Nurse	256	Pamphlet
16	Village health worker	512	Church

Other (specify)...

....

8. Na u se kile oa bona lipampiri tsee?
(*Circle one answer)

Yes No

9. Name of Respondent:_______________________________

10. LINE NO. FROM FORM A:_________

11. *Batho ba bang ba neng ba le teng ha foromo e tlatsoa.
(Ba ka 'na ba feta bo ngoe)

1	Husband
2	Wife
4	Other males
8	Other females
16	Children
32	No-one

APPENDIX F

EXAMPLE SURVEY QUESTIONNAIRE TO DEMONSTRATE THE USE OF A PALMTOP COMPUTER AS DISCUSSED IN CHAPTER 3

QUESTIONNAIRE NUMBER □

ROUND □

PSU □□□□□

HOUSEHOLD NUMBER □□

PERSON NUMBER □□

Mark the appropriate block with a 'X' where applicable.

1) SURNAME: ______________________________

2) FIRST NAMES: ____

3) SEX:

Male	Female

4) AGE (in completed years): _______

MINING SECTION

(To be asked of males over 15 years)

5) EVER WORKED ON MINES

Not Applicable	1
Never worked on mines	2
Worked on mines but never worked underground	3
Less than five years underground	4
5 to 10 years underground	5
More than ten years underground	6

MATERNAL CARE SECTION

(To be askeu ɔf females aged 15-49 years)

6a) ARE YOU PREGNANT ?

Yes	No	Don't Know

(If 'NO' then go to Question 7a)

6b) WHERE DO YOU INTEND TO HAVE THE BABY DELIVERED ?

Hospital	1
Clinic (health centre, centre, dispensary)	2
Private doctor/nurses' surgery	3
Home	4
Other	5

If 'Other' specify ______________________________

(If 'Home' or 'Other' in Question 6b) then go to Question 6c) else go to Question 7a))

6c) WHO WILL ASSIST IN THE DELIVERY?

Mopepisi	1
Relative	2
Friend	3
Other	4

If 'Other' specify ______________________________

6d) There are many reasons why women do not deliver in a health facility ? Why will you not?

Lack of money	1
Facilities too far	2
The birth came unexpecteoly	3
Clinic was closed	4
Prefer to deliver at home	5
Do not find it necessary to deliver at clinic	6
Attitude of clinic/hospital staff	7
Tradition is to deliver at home to learn how to deliver others	8
Delivered at home so that could bury placenta	9
Other	10

If 'Other' spec fy ______________________________

INJURY SECTION

7a) WHAT INJURIES DO YOU HAVE ?

Broken bone	1
Torn skin or flesh	2
Burns	3
Head Injury	4
Sprain or Bruise	5
Eye injury	6
Other	7

If 'Other' specify ______________________________

7b) HOW WAS THE INJURY CAUSED ?

Traffic accident	1
Domestic accident	2
Mine accident	3
Other work/farming accident	4
Fight/assault	5
Horse	6
Playing sport	7
Other	8

If 'Other' specify ______________________________

APPENDIX G

DATA ENTRY PROGRAM OF A PALMTOP COMPUTER (PSION WORKABOUT(1995)) WHICH PERFORMS VARIOUS DATA ENTRY CHECKS ON THE RESPONSES FROM APPENDIX F, AS DISCUSSED IN CHAPTER 3

```
* This program performs various checks on the data input for
* the questionnaire in Appendix F, and display a data entry
* error when necessary. It ensures that only the necessary
* questions appear on the screen. For example for males, the
* Maternal care questions will not appear on the screen. It
* also ensures that the linking variables are complete, so
* that merging of the data for further analysis is made
* simple. Finally the data is saved in an ASCII data file.
```

```
PROC wits:
        local k%,n$(128),f$(128),sex%,age&,height
        local m1%,m2&,m3&
        local p1%,p2%,p3%,p4%,p2$(128),p3$(128),p4$(128)
        local i1%,i2%
        local mine%,mine1%,preg%,clin%,inj%,sun%
        REM file management...
        if exist("m:\dat\stats.dbf")
                open
"m:\dat\stats.dbf",a,n$,f$,sex$,height$,mine$,m2$,mine1$,m3$,preg$,p1$,p2$,p2$,p3$,p3$,p4$,clin$,p4$,inj
$,i1$,i2$,sun$
        else
                create
"m:\dat\stats.dbf",a,n$,f$,sex$,height$,mine$,m2$,mine1$,m3$,preg$,p1$,p2$,p2$,p3$,p3$,p4$,clin$,p4$,inj
$,i1$,i2$,sun$
        endif
        gIprint "Records:"+gen$(count,3)
        demo::
        sex%=1
        m1%=1
        p1%=1
        n$="" :f$="" :age&=0 :height=0 :p2$=""
        p3$="" :p4$=""
        mine%=0 :mine1%=0 :preg%=0 :clin%=0 :inj%=0 :sun%=0
        demo1::
        dInit
        dText "","Demographics:",$302
        dEdit n$,"Surname:",20
        dEdit f$,"First names:",20
        dChoice sex%,"Sex:","Male,Female"
        dLong age&,"Age (yrs):",0,140
        dFloat height,"Height (m):",0,2.5
        if dialog=0 :stop :endif
        if n$="" or f$="" or age&=0 or height=0
                beep 5,300
                gIprint "Must complete all entries!"
                goto demo1::
        endif
        if sex%=1       REM male
                if age&>15 mine::
                        dInit
                        dText "","Mining:",$302
                        dText "","Ever worked on mines?",2
                        dButtons "Yes",%Y,"No",%N
                        k%=dialog
                        mine%=k%
                        if k%=0
                                goto demo1::
                        elseif k%=%y
                                mine1::
                                dInit
                                dText "","Mining:",$302
                                dLong m2&,"Total years mining:",0,age&
                                if dialog=0 :goto mine:: :endif
                                mine2::
                                dInit
                                dText "","Mining:",$302
                                dText "","Ever worked underground?",2
                                dButtons "Yes",%Y,"No",%N
                                k%=dialog
                                mine1%=k%
                                if k%=0
                                        goto mine1::
                                elseif k%=%y
                                        mine3::
                                        dInit
                                        dText "","Mining:",$302
                                        dLong m3&,"Total years underground:",0,m2&
                                        if dialog=0 :goto mine2:: :endif
                                endif
                                dInit
                                dText "","Mining:",$302
                                dText "Years above ground:",gen$(m2&-m3&,3),2
                                dText "Years underground:",gen$(m3&,3),$202
                                dButtons "Yes",%Y,"No",%N
                                k%=dialog
                                if k%=0
                                        goto mine3::
                                elseif k%=%n
                                        goto mine::
                                endif
                        endif
                endif
        elseif sex%=2  REM female
                if age&>12 preg::
                        dInit
                        dText "","Pregnancy:",$302
                        dText "","Are you pregnant?",2
                        dButtons "Yes",%Y,"No",%N
                        k%=dialog
                        preg%=k%
                        if k%=0
                                goto demo1::
                        elseif k%=%y
                                preg1::
                                dInit
                                dText "","Pregnancy:",$302
                                dText "","Where do you intend",2
                                dText "","to have the baby delivered?",$202
                                dChoice p1%,"Options:","1. Hospital,2. Clinic,3. Private surgery,4.
Home,5. Other"
                                k%=dialog
                                if k%=0 :goto preg:: :endif
                                if p1%=5      REM other
                                        preg2::
                                        dInit
                                        dText "","Pregnancy:",$302
                                        dText "","Please specify where",2
                                        dText "","baby will be delivered:",$202
                                        dEdit p2$,"Where:",20
                                        if dialog=0 :goto preg1:: :endif
                                endif
                                if p1%=4 or p1%=5    REM home or other
                                        preg3::
                                        dInit
                                        dText "","Pregnancy:",$302
                                        dText "","Who will assist",2
                                        dText "","in the delivery?",$202
                                dChoice p2%,"Options:","1. Mopepisi,2. Relative,3. Friend,4. Other"
                                        if dialog=0 :goto preg2:: :endif
                                        if p2%=4       REM other
                                                preg4::
                                                dInit
                                                dText "","Pregnancy:",$302
                                                dText "","Please specify who will",2
                                                dText "","assist in the delivery:",$202
                                                dEdit p3$,"Who:",20
                                                if dialog=0 :goto preg3:: :endif
                                        endif
                                        preg5::
                                        dInit
                                        dText "","Pregnancy:",$302
                                        dText "","There are many reasons why",2
                                        dText "","women do not deliver in a health",2
                                        dText "","facility. Why will you not?",$202
                                        dChoice p3%,"","1. Lack of money,2. Facilities too far,3.
Birth came unexpectedly,4. Clinic was closed,5. Prefer to deliver at home,6. Not necessary to deliver at
clinic,7. Attitude of staff,8. Tradition to deliver at home,9. Other"
                                        if dialog=0 :goto preg4:: :endif
                                        if p3%=9       REM other
                                        preg6::                               dInit
                                        dText "","Pregnancy:",$302
                                        dText "","Please specify your reason for",2
                                        dText "","delivery not at health facility:",$202
                                        dEdit p4$,"Reason:",20
                                        if dialog=0 :goto preg5:: :endif
                                        endif
                                        preg7::
                                        dInit
                                        dText "","Will you and the child visit",2
                                        dText "","any clinic or hospital after birth?",2
                                        dButtons "Yes",%Y,"No",%N,"Don't know",%D
                                        k%=dialog
                                        clin%=k%
                                        if k%=0
                                                goto preg6::
                                        elseif k%=%n or k%=%d
                                                dInit
                                                dText "","Pregnancy:",$302
                                                dText "","There are many reasons why women",2
                                                dText "","do not go to a clinic during the 2",2
                                                dText "","months after birth. Why will you not?",$202
                                                dChoice p4%,"","1. Did not feel sick,2. No money,3.
Child passed away,4. Child received 1st immunisation,5. Too young to visit clinic,6. Mother was ill,7.
No reason"
                                                if dialog=0 :goto preg7:: :endif
                                        endif
                                endif
                        endif
                endif
        endif
        inj::
        dInit
        dText "","Injuries:",$302
        dText "","Do you have any injuries?",$202
        dButtons "Yes",%Y,"No",%N
        k%=dialog
        inj%=k%
        if k%=0
                if ((age&>12) and (sex%=2))
                        goto preg::
                else
                        goto demo1::
                endif
        elseif k%=%y
                inj1::
                dInit
                dText "","Injuries:",$302
                dText "","What injuries do you have?",$202
                dChoice i1%,"Options:","1. Broken bone,2. Sprain or bruise,3. Torn skin or flesh,4. Eye
injury,5. Burns,6. Head injury,7. Other"
                if dialog=0 :goto inj:: :endif
                inj2::
                dInit
                dText "","Injuries:",$302
                dText "","How was the injury caused?",$202
                dChoice i2%,"Options:","1. Traffic accident,2. Domestic accident,3. Mine accident,4.
Other work/farming accident,5. Fight/assult,6. Horse,7. Playing/sport, 8. Other"
                if dialog=0 :goto inj1:: :endif
                dInit
                dText "","Injuries:",$302
                dText "","Did this happen since",2
                dText "","the previous Sunday?".$202
                dButtons "Yes",%Y,"No",%N
                k%=dialog
                sun%=k%
                if k%=0 :goto inj2:: :endif
        endif
        dInit
        dText "","All complete!",$302
        dButtons "Continue",27
        dialog
        a.n$=chr$(34)+"Surname:"+n$+chr$(34)
        a.f$=chr$(34)+"First names:"+f$+chr$(34)
        if sex%=1
                a.sex$=chr$(34)+"Sex:Male"+chr$(34)
        else
                a.sex$=chr$(34)+"Sex:female"+chr$(34)
        endif
        a.height$=chr$(34)+"Height:"+gen$(height,5)+"m"+chr$(34)
        if mine%=%y
                a.mine$=chr$(34)+"Worked on mines"+chr$(34)
        else
                a.mine$=chr$(34)+"Never worked on mines"+chr$(34)
        endif
        a.m2$=chr$(34)+"Total years on mine:"+gen$(m2&,3)+chr$(34)
        if mine1%=%y
                a.mine1$=chr$(34)+"Worked underground"+chr$(34)
        else
                a.mine1$=chr$(34)+"Never worked underground"+chr$(34)
        endif
        a.m3$=chr$(34)+"Total years underground:"+gen$(m3&,3)+chr$(34)
        if preg%=%y
                a.preg$=chr$(34)+"Pregnant"+chr$(34)
        else
                a.preg$=chr$(34)+"Not pregnant"+chr$(34)
        endif
        a.p1$=chr$(34)+"Delivery choice:"+gen$(p1%,3)+chr$(34)
        a.p2$=chr$(34)+p2$+chr$(34)
        a.p2$=chr$(34)+"Assistance choice:"+gen$(p2%,3)+chr$(34)
        a.p3$=chr$(34)+p3$+chr$(34)
        a.p3$=chr$(34)+"No clinic choice:"+gen$(p3%,3)+chr$(34)
        a.p4$=chr$(34)+p4$+chr$(34)
        if clin%=%y
                a.clin$=chr$(34)+"Will visit after birth"+chr$(34)
        else
                a.clin$=chr$(34)+"Will not visit after birth"+chr$(34)
        endif
        a.p4$=chr$(34)+"Reason choice:"+gen$(p4%,3)+chr$(34)
        if inj%=%y
                a.inj$=chr$(34)+"Injury"+chr$(34)
        else
                a.inj$=chr$(34)+"No injury"+chr$(34)
        endif
        a.i1$=chr$(34)+"Type of injury:"+gen$(i1%,3)+chr$(34)
        a.i2$=chr$(34)+"How caused:"+gen$(i2%,3)+chr$(34)
        if sun%=%y
                a.sun$=chr$(34)+"Happened since Sunday"+chr$(34)
        else
                a.sun$=chr$(34)+"Did not happened since Sunday"+chr$(34)
        endif
        append
        gIprint "Added!"
        goto demo::
ENDP
```

APPENDIX H

AN IMPLICIT DIFFERENTIATION METHOD FOR ESTIMATING THE VARIANCE FOR A VECTOR OF SURVEY STATISTICS (FROM CHAPTER 8 SECTION 8.4)

Binder (1981, 1983) proposed and justified using an implicit differentiation method for estimating the covariance matrix for a vector of survey statistics. Logistic regression coefficients fall into this category of parameters that are implicitly defined.

Let $\beta = (\beta_0, \beta_1, \ldots, \beta_q)'$ be the population parameter vector in the model:

$$W(\beta) = \sum_{k=1}^{N} W(Z_k; \beta)$$

where N is the number of observations in the population, $Z_k = (z_{1k}, \ldots, z_{qk})$ are the data values for the k-th unit, and $W(\beta)$ is a vector with i-th element

$$W_i(\beta) = \sum_{k=1}^{N} W_i(Z_k; \beta)$$

which is estimated from the sample by $\hat{W}(\beta)$. $\hat{W}(\beta)$ is the estimator based on the functions of data values $W(Z_1;\beta), \ldots, W(Z_N;\beta)$.

Assuming that a unique solution exists, then $\hat{\beta}$, the maximum likelihood estimate of β, is defined as the solution to $\hat{W}(\beta)=0$.

To approximate the variance of $\hat{\beta}$ Binder expands $\hat{W}(\hat{\beta})$ in a Taylor series about the point $\hat{\beta} = \beta$, where β is the true unknown parameter. Defining $\hat{J}(\beta) = \partial\hat{W}(\beta)/\partial\beta$ as the q x q matrix whose ijth element is the partial derivative $\partial\hat{W}_i(\beta)/\partial\beta_j$, and expanding $\hat{W}(\beta)$ about $\hat{\beta} = \beta$ using the first two terms of the Taylor expansion gives:

$$0 = \hat{W}(\hat{\beta}) \approx \hat{W}(\beta) + \hat{J}(\beta)(\hat{\beta}-\beta)$$

or if $\hat{J}^{-1}(\beta)$ exists

$$\hat{\beta}-\beta \approx -\hat{J}^{-1}(\beta)\hat{W}(\beta)$$

This leads to the approximation of the covariance matrix of $\hat{\beta}$:

$$\begin{aligned} V(\hat{\beta}) &\doteq E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \hat{J}^{-1}(\beta)E[\hat{W}(\beta)\hat{W}'(\beta)][\hat{J}^{-1}(\beta)]' \\ &= \hat{J}^{-1}(\beta)[V(\hat{W}(\beta))][\hat{J}^{-1}(\beta)]' \end{aligned}$$

where $V(\hat{W}(\beta))$ is the covariance matrix of $\hat{W}(\beta)$. Finally, β is replaced by its estimator $\hat{\beta}$, in both $\hat{J}^{-1}(\beta)$ and $V(\hat{W}(\beta))$ to obtain the estimator of the covariance matrix of $\hat{\beta}$:

$$\hat{V}(\hat{\beta}) = [\hat{J}^{-1}(\hat{\beta})][\hat{V}(\hat{W}(\hat{\beta})][\hat{J}^{-1}(\hat{\beta})]'$$

APPENDIX I

SAS PROGRAM TO IDENTIFY OUTLYING OBSERVATIONS FOR CHILD NUTRITION DATA, AS DISCUSSED IN CHAPTER 5

```
data sss;set dat.int; one=1; keep waz haz whz one;
* PROC IML is used to perform the matrix operations

proc iml;
use sss var {waz haz whz one};
read all var {waz haz whz one};
show names;
read all var {waz haz whz} into x ;
n=nrow(x);
xpxi=(t(x)*x);
m=(t(x)*one*t(one)*x)/n;
covm=(xpxi-m)/n;
mean=(t(x)*one)/n;

* The number of records to be tested for outliers is specified
* and the records lying in the tail areas of a chi-square
* distribution are identified

do i=1 to 9943;
a=1:3;
m=x[i,a];
chisq=(m-t(mean))*inv(covm)*t(m-t(mean));
prchisq=1-probchi(chisq,0.05,3);

* Finally the valid records are written to a file

filename out 'dat.clean';
file out;
put prchisq;
end;
data qqq;
infile '~/dat.clean';
input  prchisq;
if prchisq < 0.05;
proc print;run;
```

APPENDIX J

SAS PROGRAM TO CALCULATE THE PEARSON CHI-SQUARE GOODNESS-OF-FIT TEST FOR A LOGISTIC REGRESSION MODEL - CHAPTER 8, SECTION 8.8.2

```
* This SAS program calculates X2, the Pearson chi-square test
* statistic to test the goodness-of-fit of the fitted model
* when there are discrete predictor variables in the
* logistic regression model

* The predicted values from the SUDAAN analysis are read into
* a SAS data set

goptions device=a4paper;
*options device=tek4014;
data ggg;
infile '~/g24.dbs';
input xxx expect y resid sex age sfed educ round source;
data g2;set ggg;

proc sort ; by sfed educ round source;

proc means; var  y;
 by sfed educ round source;
output out=mjp sum= ysum;
proc print data=mjp;run;
run;

* The expected and predicted numbers in each combination   *
* of the predictor variables are calculated                *

data www;set g2 ;
 by sfed educ round source;
if first.sfed  or first.educ  or first.round
or first.source;
keep sfed  educ round source expect;
run;

* The standardised residuals and the overall chi-square test *
* statistic are calculated                                    *

data qqq;set mjp; nn=_freq_;
proc sort;
 by sfed educ round source;
data sss; merge www qqq;
 by sfed educ round source;
stdres=(ysum - nn*expect)/(nn*expect*(1-expect))**0.5;
chisqt=(ysum - nn*expect)**2/(nn*expect*(1-expect));
proc print;run;
proc means;
var chisqt;
output out=chi sum=chisum;
run;
proc print data=chi;run;
```

APPENDIX K

SAS PROGRAM TO CALCULATE THE HOSMER LEMESHOW GOODNESS-OF-FIT TEST FOR A LOGISTIC REGRESSION MODEL - CHAPTER 8, SECTION 8.8.3

```
* This SAS program calculates Ĉ, the Hosmer Lemeshow test
* statistic to test the goodness-of-fit of the fitted model
* when there are continuous predictor variables in the
* logistic regression model


* The predicted values from the SUDAAN analysis are read into
* a SAS data set


data ggg;
infile '~/124b.dbs';
input xxx expect y resid sex age sfed educ round source;
data ddd;set ggg; if expect>0;if age<24;
proc sort;by sex age sfed educ source;
proc sort ; by sfed sex age  educ source ;


* The 10 groups are created according to the predicted values

data g1;set ddd;
if expect < 0.00293    then group=1;
if expect >= 0.00293    and expect < 0.024442 then group=2;
if expect >= 0.024442   and expect < 0.059508 then group=3;
if expect >= 0.059508   and expect < 0.076747 then group=4;
if expect >= 0.076747   and expect < 0.093827 then group=5;
if expect >= 0.093827   and expect < 0.11412  then group=6;
if expect >= 0.11412    and expect < 0.13296  then group=7;
if expect >= 0.13296    and expect < 0.16100  then group=8;
if expect >= 0.16100    and expect < 0.20060  then group=9;
if expect >= 0.20060    then group=10;
proc print;run;
proc sort ; by group sfed sex age educ source ;



* The observed count in each group is calculated

proc means noprint;
var  y ;
by group sfed sex age educ source;
output out=mjp sum= sumy ;



data sss;set mjp;
proc sort; by group;
proc means noprint;var sumy ;by group;
output out=mjp2 sum= sumy2  ;
```

```
data sss2;set g1;
proc sort;by group;
proc means noprint;var y; by group;
output out=mjp3 n=num ;


* The Ĉ statistic is calculated

data g1b;set g1;
proc sort;by group;
proc means; var expect; by group;
output out=fin mean=pikmn;
run;

data kkk2;merge fin mjp2 mjp3;by group;
c=((sumy2-num*pikmn)**2)/(num*pikmn*(1-pikmn));
proc print;run;
proc means noprint ; var c; by group;
output out=final sum=sumc;
proc print data=final;run;
proc means data=final;
var sumc;
output out=fil sum=smc;
proc print data=fil;run;
quit;
```

REFERENCES

AGRESTI, A. (1990). Categorical Data Analysis, New York:Wiley.

ANDERSON, M.A. (1979). Comparison of Anthropometric Measures of Nutritional Status in Preschool Children in Five Developing Countries The American Journal of Clinical Nutrition, 32(11):2339-2345.

BAIRAGI, R. ET AL. (1983). The influence of nutritional status on age misstatement for young children in rural Bangladesh, Cornell Nutritional Surveillance Programme Working Paper Series, July 1983, No. 27 . Ithaca New York.

BAIRAGI, R. (1983). On error in the estimate of malnutrition due to bias and random error in anthropometry and age. Cornell Nutritional Surveillance Programme Working Paper Series, June 1983, No. 17. Ithaca, New York.

BENNET S., WOODS A.J., LIYANAGE, and SMITH D.L. (1991) A simplified general method for cluster sample surveys of health in developing countries. World Health Statistics Quarterly; 44:98-106. Medicine, 1989.

BINDER, D.A.(1981), On the Variance of Asymptotically Normal Estimators from Complex Surveys, Survey Methodology, Vol. 7.

BINDER, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys, International Statistical Review, 51, pp. 279-292.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W., (1975) Discrete Multivariate Analysis: Theory and Practice. Cambridge: MIT Press.

CAMERON, N. (1991). Measurement Issues Related to Anthropometric Assessment of Nutritional Status, In: Anthropometric Assessment of Nutritional Status ed. John Himes, pp. 347-364.

CANTOR, D.C., ROJAS, G.,(1991). Future Prospects for Survey Processing in Developing Countries: Technologies for Data Collection, Demographic and Health Surveys World Conference, Vol. 2.

CLOGG, C.C. and ELIASON, S.R.(1987). Some common problems in log-linear analysis. Sociological Methods and Research 16, p.8-44.

COCHRAN, W.G.(1977). Sampling Techniques, Wiley and Sons.

CROFT, T.,(1991). DHS Data Editing and Imputation, Demographic and Health Surveys World Conference, Vol. 2.

CUSHING, J. (1991). DHS Data Processing Strategy: Advantages and Disadvantages, Demographic and Health Surveys World Conference, Vol. 2.

DAVIES, D.P. In: Waterlow J.C., ed. (1988). Linear Growth Retardation in Less Developed Countries, Nestle' Nutrition Workshop series. Vol. 14. New York: Nestle' Nutrition Vevey: Raven Press.

DIBLEY, M.J., GOLDSBY, J.B., STAEHLING, N.W., and TROWBRIDGE, F.L. (1987a). Development of Normalized Curves for the International Growth Reference: Historical and Technical Considerations, American Journal of Clinical Nutrition 46(5):736-748.

DIBLEY, M.J., STAEHLING, N., NIEBURG, P., and TROWBRIDGE, F.L. (1987b). Interpretation of Z-score anthropometric indicators derived from the international growth reference, American Journal of Clinical Nutrition, 46:749-62.

EFRON, B. (1982). The Jacknife, the Bootstrap and other resampling plans, Society for industrial and Applied Mathematics, Philadelphia, Pensylvania, U.S.A.

FAY, R.E., (1983), On Adjusting the Pearson Chi-Square Statistic for Cluster Sampling, in Proceedings of the Social Statistics Section, American Statistical Association, pp. 665-670.

FAY, R.E., (1983), CPLX-Contingency Table Analysis for Complex Sample Designs, Program Documentation, unpublished report, U.S. Bureau of the Census.

FELLEGI, I.P.(1980), Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples, Journal of the American Statistical Association, 75, 261-268.

FERRINHO, P., VALLI, A., GROENEVELD, T., BUCH E., and COETZEE, D. (1992), The effects of cluster sampling in an African urban setting, Central African Journal of Medicine, Vol. 38, No. 8.

FOLSOM, R.E. (1974). National Assessment Approach to Sampling Error Estimation, Sampling Error Monograph. Prepared for National Assessment of Educational Progress, Denver, CO.

GORSTEIN, J. (1989). Assessment of Nutritional Status: effects of different methods to determine age on the classification of undernutrition, Bulletin of the WHO, 67(2):143-150.

GRAITCER, P.L., GENTRY, E.M. (1981). Measuring children: One reference for all, Child Health, The Lancet.

GRIZZLE, J.E., STARMER, C.F., and KOCH, G.G. (1969). Analysis of Categorical Data by Linear Models. Biometrics, 25, pp. 489-504.

HAAGA, J.G. (1986). Negative Bias in Estimates of the Correlation between Children's Weight for Height and Height for Age, Growth, 50: 147-154.

HARTLEY, H.O., and ROSS, A. (1954). Unbiased ratio estimates. Nature, 174, 270-271.

HENDERSON, R.H.,and SUNDARESAN, T. (1982). Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. Bulletin of the World Health Organization, 60 (2):253-260 (1982).

HOSMER, D.W., and LEMESHOW, S. (1989). Applied Logistic Regression, Wiley and Sons.

HUBER, P.J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1:221-233.

HUSS-ASHMORE, R. and GOODMAN, J.L. (1989). Seasonality of Work, Weight, and Body Composition for Women in Highland Lesotho, Working Paper for Department of Anthropology, University Museum, Philadelphia, USA.

JOHNSTON, F.E. (1981). Anthropometry and nutritional status, In: Assessing Changing Food Consumption Patterns, Committee on Food Consumption Patterns. Food and Nutrition Board, National Research Council. National Academy Press, Washington D.C., pp. 252-264.

KALTON, G (1983). Compensating for Missing Survey Data. Research Report Series. Ann Arbor, MI: Institute for Social Research, University of Michigan.

KALTON, G. and KASPRZYCK D. (1982). "Imputing For Missing Survey Response." American Statistical Association Proceedings of the Section on Survey Research Methods, 22-31.

KASS, G.V., (1980). Significance testing in automatic interaction detection. Appl.Statist., 24, 178-189.

KELLER, W., (1991). Anthropometric Assessment of Nutritional Status, p113-122, Wiley-Liss, Inc.

KENYA, P.R. (1990). Surveillance Methodology for planning, and evaluation of nutritional status, East African Medical Journal.

KISH, L.(1965). Survey Sampling. New York:Wiley.

KISH, L., GROVES, R.M., and KROTKI, K.P. (1976). Sampling Errors for Fertility Surveys. WFS Occasional Papers no. 17.