NON-RESPONSE ERROR IN SURVEYS

by

MONICA TALJAARD

submitted in part fulfilment of the requirements for
the degree of

MASTER OF COMMERCE

in the subject

STATISTICS

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: MRS R J EISELEN
JOINT SUPERVISOR: PROF D M SCHULTZ

JUNE 1997

SUMMARY

Non-response is an error common to most surveys. In this dissertation, the error of non-response is described in terms of its sources and its contribution to the Mean Square Error of survey estimates. Various response and completion rates are defined. Techniques are examined that can be used to identify the extent of non-response bias in surveys. Methods to identify auxiliary variables for use in non-response adjustment procedures are described. Strategies for dealing with non-response are classified into two types, namely *preventive* strategies and *post hoc* adjustments of data. Preventive strategies discussed include the use of call-backs and follow-ups and the selection of a probability sub-sample of non-respondents for intensive follow-ups. Post hoc adjustments discussed include population and sample *weighting adjustments* and raking ratio estimation to compensate for unit non-response as well as various *imputation* methods to compensate for item non-response.

Key terms:

Non-response bias; Unit non-response; Item non-response; Response rate; Completion rate; Substitution; Call-backs and Follow-ups; Sub-sampling; Response mechanism; Sample weighting; Population weighting; Post-stratification; Raking ratio estimation; Imputation; Distance function matching

ACKNOWLEDGEMENTS

I would like to thank the following people for their help and support:

Holy Spirit, who is my constant Friend and Guide

I'm thanking you, God, out loud in the streets,
singing your praises in town and country.
Great is your love, reaching to the heavens;
your faithfulness reaches to the skies.

(Psalm 57:9-10)

My husband Des, who had to make many sacrifices but remained the encourager (Thanks for all your long-suffering trips to the Colonel's place when our kitchen just would not respond!)

Heather Herbst, who proofread my work

Gavin Ludick, who sorted out the endless hardware and software problems I experienced while typing this dissertation

Riëtte Eiselen - jy was altyd net 'n telefoonoproep weg as die paniek my wou oorval. Baie dankie vir al die hulp, ondersteuning en aanmoediging!

HIERDIE OOMBLIK

as onherhaalbaar ervaar,

na willekeur binnewaarts gaar:

dáár waar alles gedy,

vir altyd veramber bly.

Ontvang wat jou toe-kom, waak

teen afding of aanspraak maak.

Ontvang en afsien van dit

wat sink tot onsigbaar besit.

Elisabeth Eybers, <u>Versamelde Gedigte</u>

TABLE OF CONTENTS

CHAPTER 5: COMPENSATING FOR UNIT NON-RESPONSE

I have a small sample size [number of respondents], *n = 101. The population* [sample size] *was 1400 ... I did a mail survey and this is the return I have ... While I know that 7% is a very small sample size* [response rate], *what I need is any literature citations or sources where it would indicate that a sample size like this is acceptable, but given certain caveats. I know already that this sample size is small, but the research is done and out of money. If I can only draw extremely limited conclusions, fine. I just need some citation from a journal which deals with the subject and might indicate that such a sample might give insight or is limited in generalizability but still serves some purpose.*

Scuralli, J. (1996, February 5). *Small sample size: Help required.* [e-mail to Statistics and Statistical Discussion List: STAT-L], [Online]. Available e-mail: STAT-L@VM1.MCGILL.CA.

To do good sampling one must face the problem of nonresponse and not bury it.
William Edwards Deming, Some Theory of Sampling

INTRODUCTION

A probability sample will send the interviewer through mud and cold, over long distances, up decrepit stairs, to people who do not welcome an interviewer; but such cases occur only in their correct proportions. Substitutions are not permitted: the rules are ruthless.

William Edwards Deming, Some Theory of Sampling

At the end of the 19th century, the idea arose in some countries that the investigation of a *representative sample* of the population may have considerable advantages over a census. But controversy arose about how representativity can be obtained: either through randomised selection from the total population or through purposive selection with the aim to obtain a "miniature version" of the population (Särndal, Swensson & Wretman 1992:527). In his landmark 1934 document, Jerzey Neyman settled the continued controversy in favour of randomised selection methods which have the advantage of making possible "... an estimate of the accuracy of the results obtained ... irrespectively of the unknown properties of the population studied" (Neyman 1934:585). The extensive development of probability sampling methods and with it, methods to control *sampling error* followed. However, as early as 1926, important contributors to the development of sampling methods stressed the need to control multiple sources of errors which may destroy the relevance of the probability sampling formula (Lessler & Kalsbeek 1992:4). These non-sampling errors arise when certain ideal conditions for the probability sample survey do not hold. The probability sample survey ideally requires (Särndal *et al.* 1992:537):

1. The construction of a frame such that every element in the target population "appears on the list separately, once, only once and nothing else appears on the list" (Kish 1965:53)

2. The selection of sample elements with their appropriate positive probabilities as assigned by the probability sampling design

3. The measurement of the "true value" of *each* variable under investigation for *every* element in the sample

4. That data processing - including coding, editing, imputation for missing values and outlier detection and treatment - occurs without any errors

5. That *valid* statistical inferences are made, including point estimation, variance estimation and calculation of confidence intervals

When these ideal conditions hold, *sampling error* is the only error present in estimates. However, as the despairing Mr. Joe Scuralli in the epigraph to this dissertation experienced, these conditions do not always hold in practice. Non-response is just one of the many types of *non-sampling errors* that may afflict a survey. Although probability sampling provides the theory that allows one to calculate accurate measures of the sampling error in a particular survey, no theory exists that allows one to estimate non-sampling errors accurately.

Violation of the above ideal conditions for probability sampling surveys generally results in four types of (non-sampling) errors:

1. *Errors of non-observation* due to failure to measure some of the elements in the selected sample (Cochran 1977:359). This may occur because (a) the frame does not give access to all elements in the target population (under-coverage), (b) no responses are obtained from some elements in the sample (unit non-response), or (c) incomplete responses are obtained from some elements in the sample (item non-response).

2. *Measurement errors* due to failure to accurately measure the true value for one or more elements in the sample. Measurement errors may arise from four sources: (a) the interviewer, (b) the respondent, (c) the questionnaire, or (d) the mode of data collection.

3. *Processing errors* introduced in the editing, coding and tabulation of results. This includes errors in data entry and imputation errors when attempting to create a rectangular data matrix suitable for computer analysis.

4. *Errors in inference* including making inferences to domains or populations not covered by the survey and using incorrect weights to adjust for unequal inclusion probabilities and/or response probabilities.

Deming (1944) lists many other sources of errors, including failure to recognise changes that take place in the target population before the results are published, failure to understand definitions and personal bias in interpretation of survey results. In this dissertation, attention is limited to errors of non-observation, specifically non-response error: a non-sampling error which is prevalent in *most* surveys. However, for many surveys, especially in undeveloped countries or areas, problems of incomplete data due to under-coverage may be far more serious than problems due to item or unit non-response (Dempster & Rubin 1983:6).

Part I of this dissertation consists of two chapters which provide the necessary background for a study of survey non-response. The introductory Chapter 1 presents little towards an elucidation of the non-response problem, except for providing a definition of the two types of non-response that commonly appear in surveys, namely, unit non-response and item non-response. Chapter 1 is important in that it aims to establish a common vocabulary. It also deals with the elementary principles and results of survey sampling so that the remainder of the chapters can provide unimpeded (and more specific) discussions of the non-response problem.

Chapter 2 deals with three diverse issues, namely, the various *reasons* for non-response commonly encountered in surveys, the *effects* of non-response on survey estimates and the *calculation* of (non-)response rates. A categorisation of the diverse reasons why sample elements do not respond is useful for showing the separate effects that various types of non-response have on survey results and the separate treatment they require. If a survey is subject to non-response, there exists no unbiased estimators of population values unless certain model-assumptions are made. This statement is substantiated from both the deterministic and stochastic viewpoints of non-response. The bias of non-response is expressed as the product of two components: the non-response rate and the differences between respondents and non-respondents. In the last section of Chapter 2, an attempt is made to standardise the many diverse definitions of the response rate that can be found in survey research.

Part II of this dissertation focuses on various methods of dealing with the non-response problem and addresses both the design of surveys to minimise non-response and the analysis of surveys with non-response.

Chapter 3 takes an empirical approach to the study of non-response: methods are considered that can be used to *identify* the extent of non-response bias in surveys *before* corrective action is taken. This includes the identification of auxiliary variables which are correlated with response behaviour and which can be used to study differences between respondents and non-respondents, the analysis of survey results at various stages in the survey and methods that can be used to obtain numerical estimates of the non-response bias in the survey. These methods are illustrated by means of a selection of identification studies reported in literature. These studies show that there are usually systematic (non-random) differences between respondents and non-respondents in many surveys.

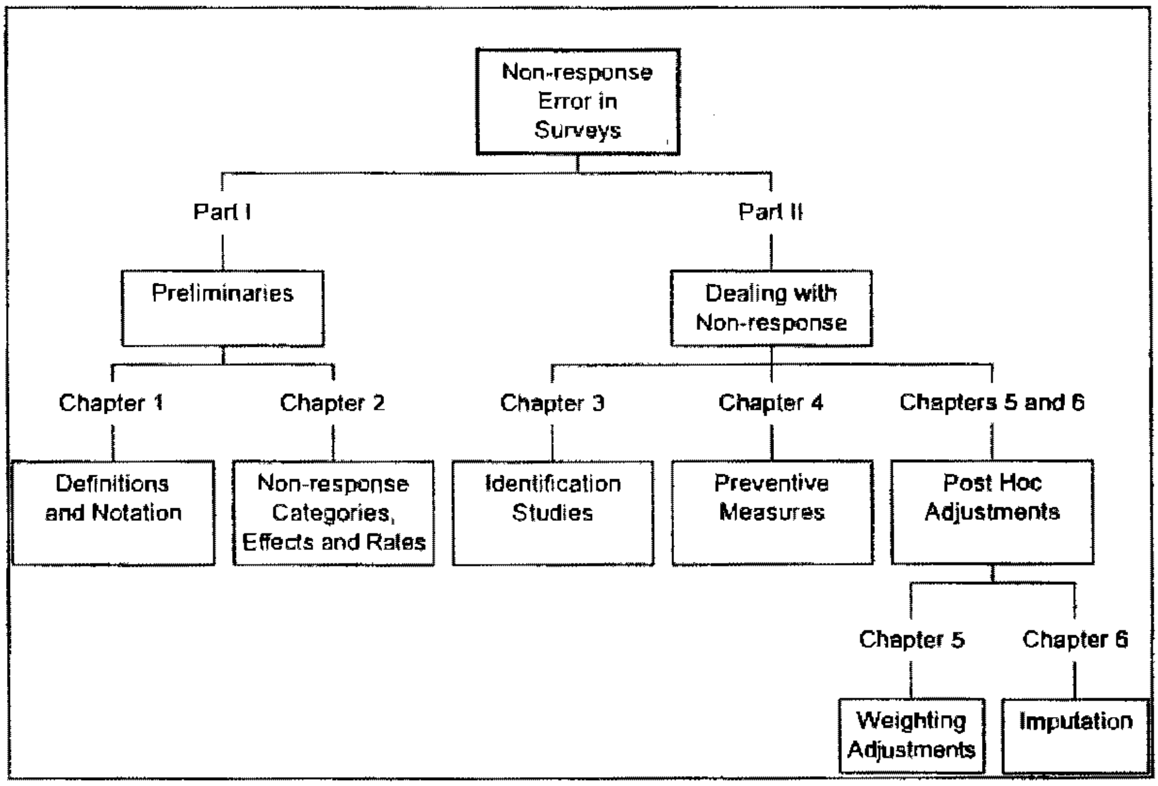
It may be advantageous to also conduct identification studies *after* adjusting for non-response to determine how effective the adjustments were and/or whether various assumptions are satisfied, for example, the assumption that the non-respondents are a random sub-sample from the corresponding adjustment class in the population.

Unfortunately, no statistical technique can be relied upon to entirely eliminate non-response. Consequently, the ideal way to handle non-response is to obtain complete data. Chapter 4 discusses various steps and data collection strategies that can be taken to prevent non-response from occurring in the first place or from becoming too large a problem. A technique which is often used but which does not lead to a reduction of the non-response problem, namely *substitution* of non-respondents with respondents, is considered briefly. This is followed by a discussion of the most successful means for reducing the survey non-response rate, namely, calling back repeatedly on non-respondents and urging them to respond. Call-backs are often restricted to a random *sub-sample* of initial non-respondents, allowing more effective use of resources to persuade sample elements to respond. This technique is discussed in the last section of Chapter 4.

In practice, even after all efforts deemed cost-effective by the surveyor have been completed, some non-response will remain. Most frequently, one has to make do with the incomplete data and *compensate* for non-response (post hoc) by adjusting the weights of survey respondents, recognising that no adjustment can fully compensate for the missing data. Chapters 5 and 6 address compensation methods for two kinds of non-response: Chapter 5 discusses *weighting* as a compensation technique for unit non-response while Chapter 6 considers *imputation* as a compensation technique for item non-response.

An outline of this dissertation is given below:

Outline of dissertation

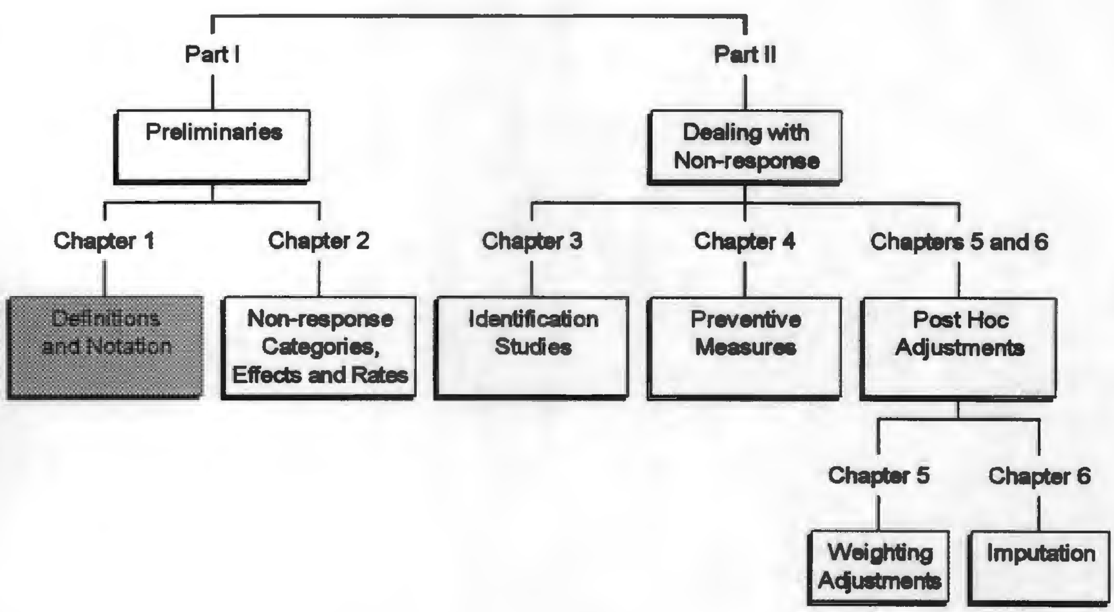


PART I

PRELIMINARIES

CHAPTER 1

DEFINITIONS AND NOTATION



CHAPTER OUTLINE

CHAPTER 1

DEFINITIONS AND NOTATION

Communication ... involves a translation from personal to public discourse ... [It] may include a description of what was observed and at least a rudimentary interpretation of what the observation means ... Communication ... depends upon the use of symbols with variable meanings and a glossing over of portions ... thought to be nonessential or mutually understood. The process of translation into verbal or written language inevitably distorts the observation.

Chadwick, Bahr & Albrecht, Social Science Research Methods

1.1. INTRODUCTION

A discussion of relevant statistical terms is an inescapable part of this dissertation. Whereas all the subsequent chapters are devoted to the particular problem of non-response in surveys, the present chapter is, of necessity, of a more tedious and elementary nature. In section 1.2, various terms and concepts which are commonly used in survey sampling theory and which are relevant to a discussion of non-response error are defined. The aim of section 1.2 is to establish a common vocabulary. In section 1.3, two important sampling strategies are discussed briefly, namely simple random sampling and stratified random sampling. Various other sample selection schemes with their relevant terminologies are also defined. In the last section of this chapter, two opposite but complementary approaches to statistical inference are outlined, namely the design-based approach and the model-based approach.

The fact that survey research "... is not itself an academic discipline, with a common language, a common set of principles for evaluating new ideas, and a well-organized professional reference group" (Groves 1989:1) has perhaps contributed to the "ignorance is bliss" attitude among many researchers making use of survey sampling (Lessler & Kalsbeek 1992:7). This attitude is manifested by researchers who (1) choose to "bury" survey errors by making gratuitous assumptions about the quality of the survey data, (2) use incorrect terminology (as in the epigraph to this dissertation!) or (3) use undefined or vaguely defined terms. Another factor contributing to the necessity of establishing a common vocabulary is the fact that the "labyrinthine methodology for incomplete-data situations" (Hartley & Hocking 1971:783) which *does* exist in a vast body of literature, has evolved through contributions from a *mélange* of researchers, ranging from statisticians and psychologists to market researchers, political scientists, agriculturists and sociologists, each using the vocabulary from their respective disciplines. (The wide range of researchers contributing to the non-response literature underlines the extent to which empirical research in many disciplines is dependent on the sample survey as major mode of data collection.)

1.2. SOME BASIC DEFINITIONS, NOTATION AND CONCEPTS

Most of the definitions in this section are obtained from or based on those in Kish (1965) and Särndal, Swensson and Wretman (1992) while the notation is based on that of Särndal *et al.* (1992) and Cochran (1977).

1.2.1. What is a Survey?

The term "survey" and the methods associated with it, are applied in various disciplines and to a wide variety of investigations, ranging from population censuses, public opinion polls, market research studies of consumer preferences, academic studies, epidemiological studies, and so forth (Babbie 1990:51). In this dissertation, the following definition of the term will be used:

Survey: A scientific investigation of specified sets of *elements* as they exist in their natural state with the aim to make quantitative inference to the aggregate set (Lessler & Kalsbeek 1992:1).

***Elements*:** The entities (individuals, institutions or physical objects) for which information is sought.

The aggregate set of elements is called the ***population***.

For the purposes of this dissertation, the following remarks apply to the above definition of the term "survey":

1. The "specified sets of elements" that are being investigated may be the entire population (a *census survey*) or a sub-set (*sample*) of the population (a ***sample survey***).

Census survey: A survey in which a complete investigation of the population is attempted.

Sample survey: A survey in which a (*probability*) *sample* of the population is studied.

2. The term "scientific investigation" is used with specific reference to probability samples. Although some may argue that non-probability samples are in some instances "scientific", studies using non-probability samples, such as quota samples, convenience samples or judgement samples are excluded from consideration.

Sample: Any sub-set of the population selected by means of a probability sampling technique[1].

Probability sample: A sample which is selected in such a way that every element in the population has a theoretically known and greater than zero probability to be included in the sample.

3. By stating that the population is studied in its "natural state", is meant that *experimental studies* are excluded. A study of on-the-job performance of employees after they have been randomly assigned to two training programs is therefore not considered a survey.

4. *Case studies* of only a few specific elements of the population where no inferences to the aggregate population are intended, are excluded from the definition of a survey. An anthropological study of contraceptive practices among Xhosas living in hostels on the East Rand is therefore not considered a survey.

1.2.2. Accessing the Population

Only *finite populations* will be considered in this dissertation. The term *universe* will therefore not be used as a synonym for population.

Finite population: A set of N elements where $N < \infty$ and N is called the size of the population[2] (Cassel, Särndal & Wretman 1977:4).

Universe: A hypothetical infinite set of elements generated by a theoretical model such as all possible rolls on a pair of dice (Kish 1965:7).

<u>Notation</u>: Suppose the N elements in the finite population can uniquely be identified by a label j where $j = 1,\ldots,N$. Denote the finite population as $U = \{1,2,\ldots,j,\ldots,N\}$.

---

[1] The term "*chunk*" has been used to distinguish between a sample selected by means of a probability sampling method and a sub-set of the population selected by means of a non-probability method (Deming 1953:14).

[2] N is assumed known for most of this dissertation. A general Horvitz-Thompson-type estimator in the case of unknown N is defined in section 1.3.1.2.

<u>*Notation*</u>: Denote a *sample* of n elements from U as $s = \{1, 2, \ldots, i, \ldots, n\}$ where s is a non-empty set such that $s \subseteq U$.

The aim of a survey is to make inferences about the finite population or ***domains*** of the population. For example, separate estimates may be required for the employed and the unemployed.

A ***domain*** is a specified part of the finite population to which inferences are to be made and which is specifically planned for when designing the sample (Kish 1965:7).

In most surveys, sampling is done from a list, called a *sampling frame* or *frame*, which identifies the elements of the population.

The ***sampling frame*** consists of materials or devices which identify, distinguish and/or allow observational access to the elements of the population. It is composed of a finite set of units (called *sampling units*) from which the probability sample is selected and it provides information that links the sampling units to population elements (Lessler & Kalsbeek 1992:44).

Sampling units in survey sampling are the elements or groups of elements on the frame from which the probability sample is selected.

A sampling frame may include physical lists (*list frames*) as well as procedures that can account for all the population elements without actually listing them. For example, in *area sampling*, the frame (*area frame*) consists of maps or aerial photographs from which the sample can be selected in several stages without obtaining a complete list of all population elements.

In ***area sampling***, the entire area containing the population is sub-divided into a number of smaller *area units* of which a probability sample is selected. Within these selected units, either a census is undertaken or a further *sub-sample* is taken (Moser & Kalton 1971:118).

A ***sub-sample*** is a sub-set of elements selected from the initial sample.

Area frame: A geographic frame consisting of *area units*; every population element belongs to a unique, identifiable area unit and the population elements can be identified after inspection of the area units (Särndal *et al.* 1992:12).

It must be possible with the aid of the frame to (1) identify and select population elements in a way that respects a given probability *sampling design* (see section 1.2.3) and (2) establish contact with selected elements, that is, an address, a telephone number, location on a map or another device for making contact must be specified in the frame (Särndal *et al.* 1992:9). Furthermore, some sampling designs require auxiliary information to improve sample selection and estimation. In such cases, the frame should also include a vector of *auxiliary information* for every sampling unit in the frame.

An ***auxiliary variable*** is any variable about which information is available prior to sampling. Ordinarily, the assumption is that the values of the auxiliary variables are known for all population elements. (Särndal *et al.* 1992:219.)

Notation: Denote an auxiliary variable as x, where the values x_j are known for all population elements $j = 1,\ldots,N$.

1.2.3. Probability Sampling Designs

Probability sampling designs have the following mathematical properties in common (Cochran 1977:9):

1. It is possible to define the set of distinct samples $\mathcal{L} = \{s_1, s_2,\ldots, s_M\}$ of fixed size n that can be obtained when the probability sampling procedure is applied to the finite population.
2. Each possible sample $s \in \mathcal{L}$ has assigned to it a known and non-zero probability of selection $P(S = s) = p(s) > 0$ where S denotes the random variable taking set values $s \in \mathcal{L}$. Equivalently, each element $j \in U$ has a known and non-zero probability π_j to be included in the sample (*inclusion probability*).
3. One of the $s \in \mathcal{L}$ is selected by a random process in which each s receives exactly the probability $p(s)$.

The function $p(\cdot)$ is called the ***sampling design*** (or the *sampling mechanism*). Only *fixed sample size designs* will be considered in this dissertation and all inferences will be conditional on the value of n.

A ***fixed (sample) size design*** is such that whenever $p(s) > 0$, the sample s will contain a fixed number of elements, say n (Särndal *et al.* 1992:38).

The ***inclusion probability*** π_j is the probability that $j \in U$ will be included in the sample s given the sampling design $p(s)$.

1.2.4. First and Second-Order Inclusion Probabilities

For a given sampling design $p(\cdot)$, define the N ***first-order inclusion probabilities***:

$$\pi_1,\ldots,\pi_j,\ldots,\pi_N$$

where π_j is obtained from $p(\cdot)$ as (Särndal *et al.* 1992:31):

$$\pi_j = P(j \in s \in \mathcal{L}) = \sum_{s \ni j} p(s) \tag{1.1}$$

The notation $s \ni j$ means that the sum is over those samples s that contain the given j.

Define the $\frac{N(N-1)}{2}$ ***second-order inclusion probabilities***:

$$\pi_{12}, \pi_{13}, \ldots, \pi_{jk}, \ldots, \pi_{N-1,N} \ , \qquad j \neq k$$

where π_{jk}, the probability that both j and k ($j, k \in U$) will be included in the sample, is obtained from $p(\cdot)$ as (Särndal *et al.* 1992:31):

$$\pi_{jk} = P(j \& k \in s \in \mathcal{S}) = \sum_{s \ni j \& k} p(s) \qquad (1.2)$$

1.2.5. Survey Variables, Population Values, Estimators

The important aim of a survey is to obtain information about unknown population characteristics (*variables*) such as age, income or number of children with the aim to make inferences about these characteristics in the population or in specified domains of the population.

Statistical inference entails making numerical estimates of finite ***population values*** such as means, totals, ratios and proportions using observed ***survey variable*** values.

The ***survey variables*** are the unknown population characteristics of interest in the survey.

Notation: Let y denote a survey variable and let y_j be the value of y for population element j. A sample of n elements is selected and the value of y is observed, in the absence of non-response, for each $j \in s$.

In practice, y is usually a vector of many components corresponding to, say, the Q items in a questionnaire, i.e., $\underline{y}' = (y_1, \ldots, y_q, \ldots, y_Q)$, but for most part of this dissertation, y will be treated as a single variable. Furthermore, according to the *design-based approach* (or the *model-assisted design-based approach*) y is assumed fixed and not a random variable. If a *model-based* approach to statistical inference is followed, y is considered to be a random variable. (See section 1.4.)

The ***population value*** is a numerical expression that summarises the values of some survey variable for all $j \in U$. It is the value that would be obtained if the entire population - rather than just a sample - were observed under the actual survey conditions (Kish 1965:9).

In this dissertation, the term *parameter* will not be used as a synonym for population value. A population parameter is understood to be the value that would be obtained if the entire population were observed under the actual survey conditions when observations are not subject to *measurement errors* (see section 1.2.11) (Kish 1965:9).

<u>Notation</u>: Denote the ***population total*** of y as $Y = \sum_{j \in U} y_j = \sum_{j=1}^{N} y_j$ and the ***population mean*** of y as $\bar{Y} = \frac{Y}{N}$. Denote the ***population (element) variance*** of y as $S^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(y_i - \bar{Y}\right)^2$.

The *sample statistic* whose value is used to estimate a population value is called an ***estimator***, while specific values of the estimator are called ***(point) estimates***.

A ***(sample) statistic*** is any real-valued function of the observations in the random sample $s \in \mathcal{S}$, that is computable for any outcome s. Two examples are the sample mean and sample total. (Särndal *et al.* 1992:33.)

<u>Notation</u>: Denote the ***sample mean*** of y as $\bar{y} = \frac{1}{n}\sum_{i \in s} y_i = \frac{1}{n}\sum_{i=1}^{n} y_i$. Denote the ***sample (element) variance*** of y as $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$.

<u>Notation</u>: In general, denote an *estimator* $\hat{\tilde{Y}}$ of some population value $\tilde{Y} = \tilde{Y}(U)$ as $\tilde{y}(S)$. Denote a specific *estimate* of $\tilde{Y}$ as $\tilde{y}(s)$.

The notation $\tilde{y}(S)$ is used to indicate that the value of an estimator will vary with different sample realisations of the random variable S under a specified sampling design $p(\cdot)$. The notation $\tilde{y}(s)$ is used to indicate a specific *value* of the estimator $\tilde{y}(S)$ obtained from the sample realisation s. The notation $\tilde{Y}(U)$ is used to indicate that a population value is a function of the survey variable observed for the entire population U.

1.2.6. The Sampling Distribution, Mean and Variance of an Estimator

The ***sampling distribution*** of an estimator $\tilde{y}(S)$ is a specification of all the possible values of the estimator each with its probability of occurrence under the sampling design in use $p(\cdot)$ (Särndal *et al.* 1992:39).

The mean and variance of the sampling distribution of an estimator are important measures that are used to assess certain properties of the estimator.

The ***mean*** or ***expected value*** of an estimator $\tilde{y}(S)$ is defined as:

$$E[\tilde{y}(S)] = \sum_{s \in \mathcal{S}} p(s)\tilde{y}(s) \tag{1.3}$$

The ***variance*** of an estimator $\tilde{y}(S)$ is defined as:

$$\begin{aligned} V[\tilde{y}(S)] &= E[\tilde{y}(S) - E[\tilde{y}(S)]]^2 \\ &= \sum_{s \in \mathcal{S}} p(s)[\tilde{y}(s) - E\{\tilde{y}(S)\}]^2 \end{aligned} \tag{1.4}$$

An *estimator* of $V[\tilde{y}(S)]$ is denoted as $v[\tilde{y}(S)]$.

The terms *design expectation* and *design variance* are often used with reference to (1.3) and (1.4) since these definitions refer to the variation over all possible samples that can be obtained under the given sampling design $p(\cdot)$ (Särndal *et al.* 1992:35).

The positive square root of the variance of an estimator $\tilde{y}(S)$ is called the ***standard error*** of the estimator.

1.2.7. Properties of Estimators

Two important measures of the quality of an estimator are its *bias* and *mean square error* (*MSE*).

The ***bias*** of an estimator $\tilde{y}(S)$ of $\tilde{Y} = \tilde{Y}(U)$ is defined as:

$$Bias[\tilde{y}(S)] = E[\tilde{y}(S)] - \tilde{Y} \tag{1.5}$$

The ***mean square error*** (***MSE***) of an estimator $\tilde{y}(S)$ is defined as:

$$\begin{aligned} MSE[\tilde{y}(S)] &= E[\tilde{y}(S) - \tilde{Y}]^2 \\ &= E[\tilde{y}(S) - E\{\tilde{y}(S)\}]^2 + [E\{\tilde{y}(S)\} - \tilde{Y}]^2 \\ &= V[\tilde{y}(S)] + [Bias\{\tilde{y}(S)\}]^2 \end{aligned} \tag{1.6}$$

There may exist various estimators of a population value $\tilde{Y}$. Two criteria which are often used to decide among possible estimators are (1) absence of bias and (2) *precision* or *accuracy* (Yates 1981:135).

An estimator $\tilde{y}(S)$ is said to be ***unbiased*** for $\tilde{Y}$ if:

$$Bias[\tilde{y}(S)] = 0,$$

i.e., if $E[\tilde{y}(S)] = \tilde{Y}$.

An estimator $\tilde{y}(\boldsymbol{S})$ is said to be ***biased*** for $\tilde{Y}$ if:

$$Bias[\tilde{y}(\boldsymbol{S})] \neq 0,$$

i.e., if $E[\tilde{y}(\boldsymbol{S})] \neq \tilde{Y}$.

If $\tilde{y}(\boldsymbol{S})$ is an unbiased estimator of $\tilde{Y}$ then $MSE[\tilde{y}(\boldsymbol{S})] = V[\tilde{y}(\boldsymbol{S})]$.

The ***precision*** of an estimator is defined as $\dfrac{1}{V[\tilde{y}(\boldsymbol{S})]}$, the inverse of its variance. The ***accuracy*** of an estimator is defined as $\dfrac{1}{MSE[\tilde{y}(\boldsymbol{S})]}$, the inverse of its MSE. (Kish 1965:25.)

If an estimator is biased, *accuracy* is a better measure of the quality of the estimator than *precision* alone (Kish 1965:25).

If $\tilde{y}_1(\boldsymbol{S})$ and $\tilde{y}_2(\boldsymbol{S})$ are two unbiased estimators of $\tilde{Y}$ and the variance of $\tilde{y}_1(\boldsymbol{S})$ is less than the variance of $\tilde{y}_2(\boldsymbol{S})$, the estimator $\tilde{y}_1(\boldsymbol{S})$ is said to be ***relatively more efficient*** than $\tilde{y}_2(\boldsymbol{S})$. The ratio:

$$\frac{V[\tilde{y}_1(\boldsymbol{S})]}{V[\tilde{y}_2(\boldsymbol{S})]} \tag{1.7}$$

is used to measure the ***efficiency*** of $\tilde{y}_2(\boldsymbol{S})$ relative to $\tilde{y}_1(\boldsymbol{S})$ (Freund & Walpole 1987:337).

The efficiency of an estimator may also be used to refer to the number of sample elements required to obtain a fixed precision (Kish 1965:24).

A sample is ***economical*** if the precision per element cost is high (or the variance per element cost is low) (Kish 1965:26).

An estimator $\tilde{y}(\boldsymbol{S})$ is said to be a ***consistent estimator*** (or an ***approximately unbiased estimator***) of $\tilde{Y}$ if $Bias[\tilde{y}(\boldsymbol{S})]$ is negligible in large samples, i.e., if for any $\varepsilon > 0$:

$$\lim_{n\to\infty} P\left(\left|\tilde{y}(S) - \tilde{Y}\right| \geq \varepsilon\right) = 0 \tag{1.8}$$

1.2.8. Confidence Intervals

Statistical inference traditionally takes the form of a random interval having a stated probability $1-\alpha$ (usually near unity) of containing the unknown population value. The random interval is called a *confidence interval.* (Särndal *et al.* 1992:55.)

A (1-α)100% ***confidence interval*** for $\tilde{Y}$ can be calculated as:

$$I = \left(\tilde{y}(s) \pm z_{1-\frac{\alpha}{2}} \sqrt{v[\tilde{y}(S)]} \right) \tag{1.9}$$

where $z_{1-\frac{\alpha}{2}}$ is the constant exceeded with probability $\frac{\alpha}{2}$ by the N(0;1) random variable, i.e., $P(\tilde{Y} \in I) = 1 - \alpha$. $1 - \alpha$ is called the *confidence level* of the interval.

The confidence interval (1.9) will be a ***valid*** confidence interval, i.e., it will contain the unknown population value for an approximate proportion $1 - \alpha$ of repeated samples s drawn with the given design $p(s)$ if the following two conditions are satisfied (Särndal *et al.* 1992:166):

1. The sampling distribution of $\tilde{y}(S)$ is approximately normal with mean (expected value) $\tilde{Y}(U)$ and variance $V[\tilde{y}(S)]$
2. $v[\tilde{y}(S)]$ is an approximately unbiased (consistent) estimator of $V[\tilde{y}(S)]$

1.2.9. Types of Populations

1.2.9.1. The Target Population

The finite population, U, about which information is desired is often called the ***target population***.

It is important that the target population is well-defined in terms of (1) content, (2) units, (3) extent, and (4) time (Kish 1965:7). For example, in a survey to determine the extent of drug abuse among teenagers in Gauteng, the target population may be specified as (1) all teenagers (2) in high schools (3) in Gauteng (4) in 1996. The target population must be defined in such a way that there is no doubt about the *eligibility* of a sampled element (Cochran 1977:5).

An element j is ***eligible*** for the survey if it is a member of the target population, i.e., if $j \in U$.

The target population reflects the ideal expectation of the surveyor, but often the target population cannot practically, conveniently or economically be surveyed because, for example, a good frame cannot be constructed or some population elements are geographically out of reach or too expensive to survey (Särndal *et al.* 1992:543). This leads to the definition of three types of populations which (hopefully) "approximate" the target population, namely the frame population, the survey population and the inference population.

1.2.9.2. The Frame Population

The ***frame population*** is the set of elements listed directly as units in the frame or that can be approached through units occurring in the frame (Murthy 1983:9).

In theory, there should be a one-to-one correspondence between the frame population and the target population but, in practice, there are bound to be some differences. These differences arise in the presence of the various frame imperfections revealed in section 1.2.11. If there exist serious differences between the desired target population and the frame population, the target population may be re-defined to fit the frame. According to Kish (1965:55) this should be avoided if the "orientation of the sample would be seriously deflected from its goal, but it can be used if the result of the redefinition is trivial or preferred".

Murthy (1983:9) guards against the definition of too idealistic a target population which cannot adequately be surveyed in practice:

> *It is the task of the survey statistician to ... take steps to make the frame population reflect the target population. If this is not feasible, the definition of the target population is to be reviewed and redefined, if necessary, to avoid misunderstandings and misuses of data at a later stage.*

Closely related to, but not identical to, the frame population is the *survey population*.

1.2.9.3. The Survey Population

The ***survey population*** is the population actually covered by the survey.

The survey population is basically determined by the frame population, but further differences from the target population may arise because of non-response from an identifiable part of the target (or frame) population or because of the exclusion of parts of the target (or frame) population which are too costly or inconvenient to survey (Murthy 1983:9). For example, the geographical delineation of the target population may comprise all of the RSA, while the survey population may exclude Kwa Zulu-Natal if it is known that no responses were obtained from this province due to a postal strike or it may exclude areas which are thinly populated, awkward to reach or for any other reason expensive from a fieldwork point of view.

In the above example concerning drug use among high school teenagers, suppose the available frame excludes private schools. The *frame population* may then be defined as (1) all teenagers (2) in public high schools (3) in Gauteng (4) in 1996. Suppose furthermore, that farm schools are too expensive to survey and that, due to time constraints, no efforts can be made to obtain responses from pupils who are absent from schools on the day of the survey. The *survey population* may then be defined as (1) all teenagers (2) present in public high schools in cities and towns (3) in Gauteng (4) on 12 June 1996. (Note that the survey population will differ significantly from the target population if a high proportion among those absent from school on the day of the survey are drug users.)

In most surveys it is difficult to define the survey population exactly and if there are no significant differences between the target population and the survey population, it is more convenient to refer only to the target population (Kish 1965:7). This convention will be followed throughout this dissertation.

1.2.9.4. The Inference Population

Another type of population, the *inference population*, is described by Kish (1965:7) and Lessler and Kalsbeek (1992:40) as the population to which the results of the survey are inferred. For example, the results of the survey above may be used to predict the level of drug use in Gauteng schools by the year 2000 or the results may be inferred to "teenagers in the RSA". An alternative definition by Murthy (1983:10) is:

The ***inference population*** is the *conceptual* population to which inference can be made after processing and adjustment of data.

According to this definition, the inference population differs from the target population with respect to lost and rejected units at the processing phase and with respect to any adjustment procedures for coverage error and non-response. According to Murthy (1983:10), if effective adjustment procedures are used, the inference population can be made to approximate the target population. The validity of the inferential process is determined by the degree of correspondence between the inference population and the target population.

1.2.10. Data Collection Methods

An important decision to be made in surveys is which data collection method to use. The methods of data collection most commonly used in surveys take the form of either *personal interviews* or *self-administered questionnaires* (see Figure 1.1.)

In ***(personal) interview surveys***, interviewers ask questions verbally and record the respondents' answers.

Self-administered questionnaires are completed by the respondents themselves, either in the presence or in the absence of the "interviewer".

Both these data collection methods depend on the same type of observational instrument, namely a structured *questionnaire*. (Principles of questionnaire design are discussed by various authors, for example, Moser and Kalton (1983), Warwick and Lininger (1975) and Bailey (1987).)

1.2.10.1. Personal Interview Surveys

Personal interviews may take the form of either *face-to-face interviews* or *telephone interviews*.

In ***face-to-face interview surveys***, the interviewers ask questions and record the respondents' answers in a face-to-face encounter. The interviewers are also called *field-workers*. In ***telephone interview surveys***, interviewing is conducted over the telephone.

In ***directory sampling***, the sampling frame is a telephone directory or other list of individuals and/or institutions and their telephone numbers.

It is possible to obtain interviews from households with *unlisted* telephones through the use of *random digit dialling* (RDD) *methods*. RDD avoids listings altogether in most versions, although it can be used in conjunction with listings.

Random digit dialling (RDD) is a process for mechanically dialling, in a random fashion, from all possible combinations of the digits in a set of available telephone numbers (Bailey 1987:200).

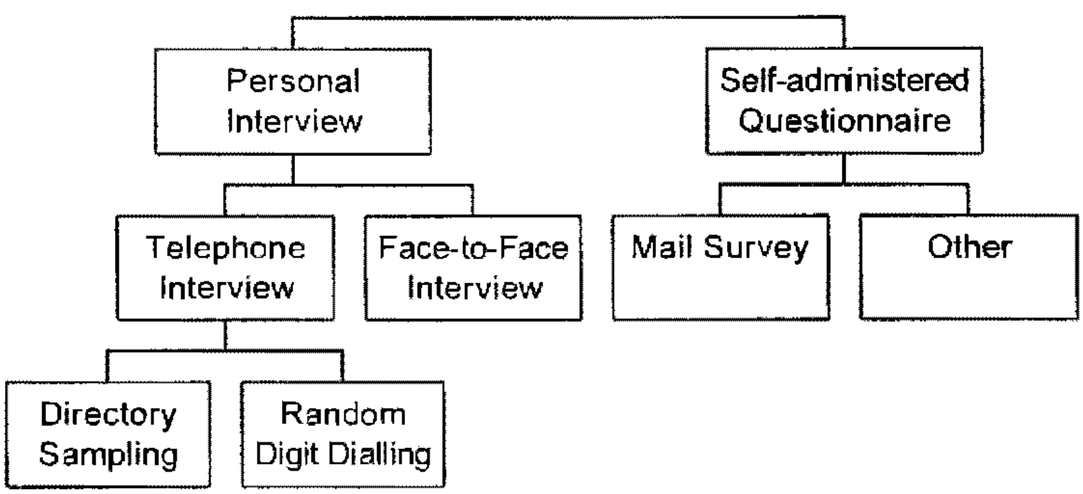
1.2.10.2. Self-administered Questionnaires

Various types of *self-administered* questionnaires may be used. For example, the questionnaire may be completed by respondents in group sessions, as in a classroom or office, or the questionnaire may be distributed in some way (in person or by mail) to the sample elements and then collected in person at a later stage, or sample elements may be requested to mail back the completed questionnaire.

> In *mail surveys*, a questionnaire, accompanied by a letter of explanation and (normally) a return envelope is mailed to each sample element. The respondents complete the questionnaires and return them to the research office by mail, using the envelope provided for the purpose. (Babbie 1990:177.)

Self-administered questionnaires have advantages and disadvantages in common with both personal interviews (face-to-face and telephone) and mail surveys. One major disadvantage of mail surveys (and some other types of self-administered questionnaires) is the low response rate that is generally obtained (see section 4.2.2). The advantages and disadvantages of each method of data collection are discussed by various authors, for example, Warwick and Lininger (1975:128), Dillman (1978) and Bailey (1987). For simplicity in this dissertation, no distinction will be made among the various types of self-administered questionnaires and the term *mail survey* will be used throughout as the major type of self-administered questionnaire.

Figure 1.1 Data collection methods



1.2.11. Sources of Errors in Surveys

1.2.11.1. Types of Survey Errors

There are two general reasons why an estimate obtained from a sample survey will deviate from its population value: (1) the estimate is calculated from data for a sub-set of the population only and (2) the observational *procedures*[3] used to produce the estimate contain imperfections or are subject to error. The deviation (*error*) due to reason (1) is called *sampling error* and errors due to a reason in (2) are called ***non-sampling errors***.

Sampling error arises because only a sub-set of the population is measured.

Non-sampling errors are all the other sources of error in a survey that are not due to sampling. Non-sampling errors are categorised as *coverage* error, ***non-response***, *measurement* errors and *processing* errors.

Coverage error refers to the difference between estimates calculated on the frame population and the same estimates calculated on the target population (Groves 1989:83).

Non-response occurs if the desired data for one or more survey variables are not obtained from one or more eligible elements in the sample.

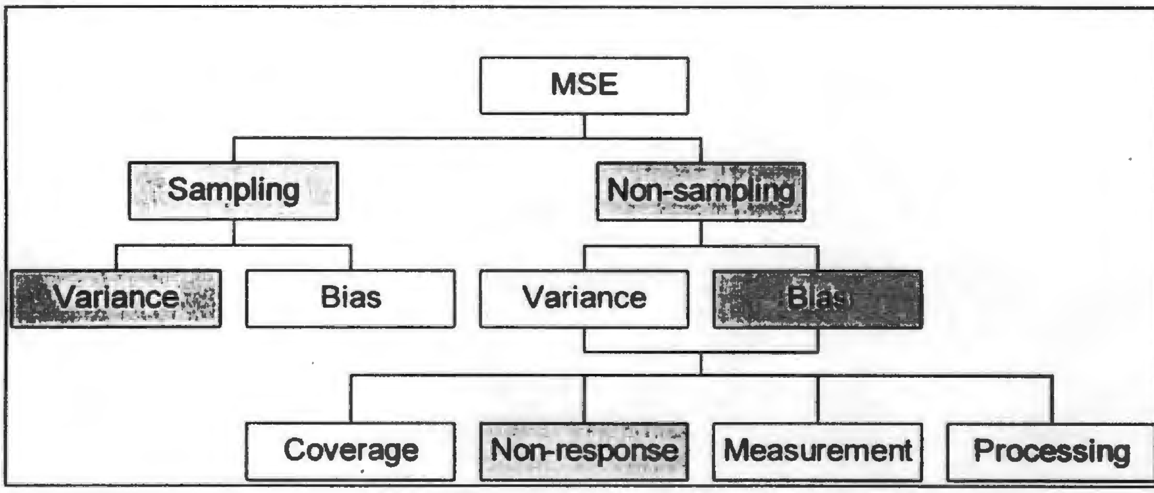
Measurement errors are errors that occur in the *data collection phase* of a survey (see section 1.2.12) when the recorded value of a survey variable for a sampled element differs from the true value. Measurement errors can be traced to four principal sources: the respondent, the interviewer, the questionnaire and the mode of data collection. (Särndal *et al.* 1992:601.)

Processing errors are errors which occur during the *data processing phase* of a survey (see section 1.2.12). These errors arise from sources such as coding, data capturing, editing, outlier treatment, imputation and tabulation. (Särndal *et al.* 1992:601.)

In equation (1.6) the MSE of an estimator (also called the *total error* of an estimator (Kish 1965:510)) was shown to consist of a variance component $V[\tilde{y}(S)]$ and a bias component $Bias[\tilde{y}(S)]$. The four-fold classification of survey errors as sampling and non-sampling *variance* and sampling and non-sampling ***bias*** is represented in Figure 1.2.

[3] This includes all the procedures that are necessary to ensure that accurate measures of the *true* value of ***each*** sample element are available at the estimation phase, for example, the construction of a frame, the measurement and the processing of data.

Figure 1.2 Sources of errors in surveys



The shaded boxes in Figure 1.2 show the various components of error that are relevant to this dissertation. In this dissertation, *sampling* error will be seen to consist mainly of a *variable* component (called *sampling variance*). Since only probability sampling designs are considered, *biases* are seen to arise mainly from *non-sampling* sources (called *non-sampling bias*), i.e., the possibility of *sampling bias* is assumed to have been eliminated by the probability sampling procedure.

The concept of *non-sampling variance*, although not considered in this dissertation, may require some clarification. All the sources of error that are normally associated with bias, namely coverage error, non-response, measurement errors and processing errors can also be associated with variable error. All that is necessary for this switch of category is that the magnitude of these errors vary over replications of the survey[4]. For example, *simple response variance* is used to denote variation in answers to the same question if repeatedly administered to the same person over different replications of the survey and *correlated response variance* is used to denote variation in responses obtained using different sets of interviewers. (Groves 1989:8.)

While variable errors can be reduced by increasing the number of "units" of some kind, i.e., either sampling units, trials or interviewers (Kish 1965:518), most

[4] The concept of variable errors inherently requires the possibility of repeating the survey with changes in some "units" over the replications (i.e., different sample elements, different interviewers, different trials). If there were no such possibility of replications, the distinction between variable errors and biases would not exist. (Groves 1989:9.)

biases cannot be reduced by increasing the size of the sample but only by improving the *quality* of some operation. *Measurement* of biases essentially depends on a method external to the survey, for example, a comparison of the survey results with previous census data.

This dissertation deals exclusively with non-sampling bias due to non-response. For convenience, the assumption is made that there are no measurement errors, coverage error or processing errors in the data.

1.2.11.2.Types of Non-response

Two types of non-response can be distinguished, namely ***unit non-response*** and *item non-response.*

Unit non-response occurs when eligible elements in the survey do not provide data on any of the survey variables or when all the provided data are unusable (Madow, Nisselson & Olkin 1983:3).

Item non-response occurs when eligible elements in the survey provide data on some, but not all, of the survey variables or the data obtained on some, but not all, survey variables are unusable (Madow, Nisselson & Olkin 1983:3).

"Element non-response" would be a more appropriate term than "unit non-response" according to the definitions provided in section 1.2.2. However, as Särndal *et al.* (1992:559) put it, the latter term is so entrenched in the minds of survey statisticians that it can hardly be changed. In this dissertation, the term "unit non-response" will therefore be used whether reference is to an *element* or a *unit* not responding.

Särndal *et al.* (1992:556) provide the following operational definition of non-response in the case of a Q-vector of survey variables:

Suppose there are Q survey variables, $y_1,\ldots,y_q,\ldots,y_Q$. These may correspond to Q *items* on a questionnaire. Let y_{jq} be the value of the variable y_q for element j. In the absence of non-response the available data consist, after the data processing phase of a complete Q-vector of observed values for every $i \in s$:

$$\underline{y}_i' = \left(y_{i1},\ldots,y_{iq},\ldots,y_{iQ}\right)$$

These n Q-vectors form a *data matrix* of dimension $n \times Q$ with no missing values. If the $n \times Q$ data matrix is incomplete, there is *non-response.*

The element i is a ***unit non-response element*** if the entire row-vector of y-values, $\underline{\mathbf{y}}_i' = (y_{i1}, \ldots, y_{iq}, \ldots, y_{iQ})$ is missing.

The element i is an ***item non-response element*** if at least one, but not all Q components of the vector $\underline{\mathbf{y}}_i' = (y_{i1}, \ldots, y_{iq}, \ldots, y_{iQ})$ is missing.

Non-response and *under-coverage* (a type of coverage error) are two types of *errors of non-observation* and, although coverage error is not discussed in this dissertation, many of the procedures used to adjust for unit non-response may also be used to adjust for under-coverage. A brief definition of the various types of coverage error is deemed necessary.

1.2.11.3. Types of Coverage Error

At least three types of coverage error can be identified, namely *under-coverage*, *over-coverage* and *duplicate listings*.

Under-coverage occurs when some elements of the target population are not in the frame population.

Over-coverage occurs when the frame population contains elements that are not in the target population.

Duplicate listings occur when some elements of the target population appear in the frame more than once.

In this dissertation, the assumption is that there is a one-to-one correspondence between the elements in the frame population and the elements in the target population, i.e., the frame is assumed perfect for the target population. However, Särndal *et al.* (1992:14) state:

> *To come up with a perfect sampling frame is not always possible in practice. Minor imperfections are often tolerated, since a perfect frame may not be obtained without excessive cost.*

It is for this reason that, in subsequent chapters, reference will occasionally be made to the presence of non-eligible elements in the frame.

In interview surveys, the eligibility status of a sampled element must often be determined in a short screening interview during data collection (see section 2.2.2). Errors may arise in this screening interview. If an element is erroneously classified as *eligible*, the additional information obtained during the survey interview may lead to a later correction of the error. An erroneous classification as *eligible* leads to over-coverage. If an element is erroneously classified as *non-eligible*, no opportunity for correcting the error necessarily occurs unless the classification is verified. An erroneous classification as *non-eligible* leads to under-coverage. (Madow, Nisselson & Olkin 1983:16.)

Almost all telephone surveys (including directory sampling and RDD methods) are subject to under-coverage because (1) telephone directories exclude households that do not have telephones and households that have unlisted phones and (2) even if RDD is used, households with unlisted telephones will be covered, but not households without telephones.

Various types of coverage error may occur in mail surveys, for example, mail may not be delivered to the sampled address (leading to under-coverage) or it may be delivered to the wrong address (leading to over-coverage). The non-return of questionnaires due to under-coverage or over-coverage may not be distinguishable from the non-return of questionnaires due to "true" non-response, unless additional information is obtained.

1.2.12. Phases of Survey Operations

1.2.12.1. Survey Planning

In the planning phase of a survey, the investigator lays the groundwork for the research project and sets the direction for subsequent activities (Lessler & Kalsbeek 1992:15). At this point, the investigator takes various steps such as stating the research objectives; delineating the scope of work; defining the target population (the target population is usually limited by amongst others, the frame that can be obtained or constructed); determining whether the available budget is sufficient to carry out the survey as planned and deciding which mode of data collection to use, with due consideration of, amongst others, the target population and the expected response rate. It is important to make provision for the reduction of non-response and measurement errors in advance, i.e., deciding how much time, labour and money are going to be spent on follow-up procedures.

1.2.12.2. Sample Selection

Before the sample can be selected, a *frame* must be constructed, usually from sources external to the survey, such as membership lists obtained from organisations, aerial photographs or census information. Two important choices that must be made before the sample can be selected are (Särndal *et al.* 1992:30):

1. The choice of a sampling design $p(\cdot)$ and a sample selection scheme that implements the sampling design
2. The choice of a suitable *estimator* for calculating an estimate of a population value of interest

The combination of a sampling design and an estimator is called a ***sampling strategy*** (Särndal *et al.* 1992:30). The specification of a suitable sampling strategy requires amongst others, decisions on the number of selection stages (see section 1.3.3.3), the type of units selected at each stage (see section 1.3.3.2), how units are to be stratified at each stage (see section 1.3.2) and the method and number of sampling units selected at each stage (Lessler & Kalsbeek 1992:20).

Once the sampling strategy has been decided upon, the sample must be selected according to a scheme that implements the specified sampling design. In some surveys the entire sample is selected before the data collection activities begin, while in other surveys the entire sample or part of the sample is selected in the field (Lessler & Kalsbeek 1992:21).

1.2.12.3. Questionnaire Construction

The construction of a survey questionnaire requires decisions on a number of issues, for example, what information is to be recorded and the format and order of questions. When designing the questionnaire, it may be useful to make provision for the collection of auxiliary information on possible non-respondents that can be used to adjust for non-response. In most interview surveys, a manual is written which contains a set of instructions for using the questionnaire and which is issued to interviewers during training (Lessler & Kalsbeek 1992:23).

Once a working draft of the survey questionnaire has been developed and a preliminary plan for collecting survey data has been worked out, the questionnaire, the interviewer instructions and various other features of the mail survey are ***pre-tested*** by using, for example, a convenience sample of elements which resemble the target population, friends or members of the research team. This is usually followed by a ***pilot study***, a small-scale replica of the main survey performed with the purpose of identifying pitfalls in the design and analysis of the sample, e.g., to determine the adequacy of the sampling frame, the suitability of the mode of data collection, the adequacy of the questionnaire, the efficiency of the interviewer instructions and the field organisation and also to obtain preliminary estimates of the variance, the expected response rate and the probable cost and duration of the final survey. (Moser & Kalton 1971:47.) A small representative sample from the population, selected in the same manner as is intended for the final survey should be used in the pilot study. The wording, format and sequence of questions used in the pilot study questionnaire or interview should be exactly the same as in the final survey. Data processing and analysis in the pilot study should occur in the precise manner intended for the final survey. (Unisa 1992:92.)

1.2.12.4. Data Collection

Before the data collection phase of the survey, interviewers are recruited and trained and data collection supervisors are hired to ensure that interviewers follow instructions. It is important that interviewers are trained in refusal conversion strategies. During data collection, it is important that regular quality control checks are carried out, for example, by selecting a representative sample of elements in the survey to check the outcome of call attempts by the interviewer.

1.2.12.5. Data Processing

Once respondents have completed the questionnaires, the next step is to manually *edit* the responses with the purpose of eliminating obvious mistakes. This is followed by the assignment of a number or letter to each response to represent the substance of the response, i.e., *coding* of the responses (Lessler & Kalsbeek 1992:27). With most computer programs for survey estimation, the (coded) data is entered in the form of a data matrix where each row of the matrix, called a ***record***, represents a respondent's set of responses and each column represents a survey variable or an ***item***. After data entry, a second phase of editing usually takes place, this time by computer.

Computer editing involves (a) checking each ***field*** (a single entry or cell in the data matrix) of every *record* to ascertain whether it contains a valid entry (whether it falls in an accepted range of values) and (b) checking entries in certain pre-determined combinations of fields to ascertain whether the entries are consistent with one another (Fellegi & Holt 1976:17). This phase also includes outlier detection and treatment and imputation for missing responses (see Chapter 6).

1.2.12.6. Estimation and Analysis

The estimation and analysis phase includes the presentation and interpretation of sample distributions and cross-tabulations of survey variables, the calculation of point estimates and confidence intervals with appropriate adjustments for non-response using suitable auxiliary information and the calculation of measures of precision. A wide variety of tools is available for making statistical inferences, for example, correlation and regression analyses. This phase also involves the computation of sampling ***weights*** (see section 1.3.3.3) which should indicate the relative likelihood of elements being included in the sample and responding. A provisional weight is often calculated as the inverse of the inclusion probability for each respondent. The provisional weight is often multiplied with a non-response adjustment factor which is the inverse of the estimated response probability (see Chapter 5). To reduce the sampling error of estimates further, the provisional weight and the non-response adjustment may be multiplied with a final post-stratification or ratio adjustment to yield the final sampling weights. (Lessler & Kalsbeek 1992:34.)

1.2.12.7. Final Documentation

A final report of the survey, containing a detailed account of all phases of the survey, is often produced. This becomes one of the principal means by which the survey can be evaluated by the research community. The final report should contain an objective report of achievements as well as an evaluation of problems experienced. (Lessler & Kalsbeek 1992:28.)

1.3. SAMPLING STRATEGIES

Two basic sampling strategies relevant to this dissertation are *simple random sampling* and *stratified random sampling*.

1.3.1. Simple Random Sampling (srs)

Simple random sampling is a basic sampling strategy which, because of its simple mathematical properties, is assumed by most statistical theories and techniques, including the statistical theory dealing with the treatment of non-response (Kish 1965:38). Simple random sampling *without replacement* (*wor*) will be assumed throughout this dissertation. However, *with replacement* (*wr*) sampling is sometimes used in practical applications to simplify formulas for variances and estimated variances of estimates calculated from *complex samples* (see section 1.3.3.3) (Cochran 1977:18).

In ***sampling without replacement***, previously selected elements cannot be re-selected, but in ***sampling with replacement***, previously selected elements are placed in the selection pool again for possible further selection (Kish 1965:37).

Simple random sampling without replacement (***srs wor***) is a method of selecting n elements out of the N such that every one of the ${}_N C_n$ distinct samples s of fixed size n has an equal selection probability of $p(s) = \frac{1}{{}_N C_n}$ (Cochran 1977:18).

In practice, an *srs* design is often implemented by drawing elements sequentially. One way to carry out simple random sampling without replacement is by means of the following sequential scheme (Särndal *et al.* 1992:26):

1. Select with equal probability $\frac{1}{N}$ a first element from the N population elements.
2. Select with equal probability, $\frac{1}{N-1}$ a second element from the remaining $N-1$ elements.
3. …

   …

n. Select with equal probability $\frac{1}{N-n+1}$ an n-th element from the $N-n+1$ elements that remain after the first $n-1$ selections.

Hence, at each of the n successive drawings, every *unselected* element has an equal probability of selection, but previously selected elements are disregarded. For example, the probability of selecting a specific element on the second draw is $\frac{1}{N-1}$ *conditional* on the probability $\frac{N-1}{N}$ that it was not selected on the first draw. Thus, the probability of selecting the element on the second draw is:

$$\frac{N-1}{N} \times \frac{1}{N-1} = \frac{1}{N}$$

In ***simple random sampling with replacement***, at any draw all N population elements are given an equal probability $\frac{1}{N}$ of being drawn, no matter how often they have already been drawn.

The first order inclusion probabilities in an *srs wor* design are obtained from (1.1) as:

$$\pi_j = \sum_{s \ni j} p(s) = \binom{N-1}{n-1} \times \left[1 \Big/ \binom{N}{n} \right] = \frac{n}{N} = f \; ; \quad j = 1, \ldots, N \qquad (1.10)$$

where $f = \frac{n}{N}$ is called the ***sampling fraction***. The second order inclusion probabilities are obtained from (1.2) as:

$$\pi_{jk} = \sum_{s \ni j \& k} p(s) = \binom{N-2}{n-2} \times \left[1 \Big/ \binom{N}{n} \right] = \frac{n(n-1)}{N(N-1)}; \quad j \neq k = 1, \ldots, N \qquad (1.11)$$

Simple random sampling without replacement is a special type of *epsem* (equal probability of selection method) sampling, because elements have the same fixed (first-order) inclusion probability of $f = \frac{n}{N}$.

In an ***equal probability of selection method (epsem)*** every population element has an equal first-order inclusion probability $\pi_j = \pi$ for all $j \in U$ (Särndal *et al.* 1992:66).

While equal first-order inclusion probabilities for each of the N elements is common to all epsem samples, *srs wor* is distinct among them because all higher-order inclusion probabilities are also equal (Kish 1965:40).

Simple random sampling is an example of *direct element sampling*, i.e., the elements are also the sampling units on the sampling frame. This differs from *cluster* *sampling*, where the sampling units are clusters containing several elements (see section 1.3.3.2).

1.3.1.1. The srs wor Estimator of the Population Total

The unbiased *srs wor* estimator of the population total $Y = \sum_{j=1}^{N} y_j$ is:

$$\hat{Y}_{srs} = N\bar{y} = N\frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.12}$$

Since $\pi_i = \frac{n}{N}$, the *srs wor* estimator of the population total may also be written as:

$$\hat{Y}_{srs} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \hat{Y}_\pi \tag{1.13}$$

where $\hat{Y}_\pi$ denotes the ***Horvitz-Thompson estimator*** (called the ***π-estimator***) of the population total. Any estimator of the population total that can be written in the form (1.13) will be called a π-estimator of the population total.

The variance of the estimator $\hat{Y}_{srs}$ is:

$$V\left(\hat{Y}_{srs}\right) = N^2(1-f)\frac{S^2}{n} \tag{1.14}$$

An unbiased variance estimator is:

$$v\left(\hat{Y}_{srs}\right) = N^2(1-f)\frac{s^2}{n} \tag{1.15}$$

1.3.1.2. The srs wor Estimator of the Population Mean

An unbiased estimator of the population mean $\bar{Y} = \frac{Y}{N}$ is obtained directly by dividing the population total estimator (1.12) by the (known) value of N:

$$\hat{\bar{Y}}_{srs} = \frac{\hat{Y}_{srs}}{N} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{1.16}$$

In general, assuming that N is known, a simple unbiased π-estimator of the population mean is:

$$\hat{\bar{Y}} = \frac{\hat{Y}_\pi}{N} \tag{1.17}$$

The variance of the estimator $\hat{\bar{Y}}_{srs}$ is:

$$V\left(\hat{\bar{Y}}_{srs}\right) = (1-f)\frac{S^2}{n} \tag{1.18}$$

with unbiased variance estimator:

$$v\left(\hat{\bar{Y}}_{srs}\right) = (1-f)\frac{s^2}{n} \tag{1.19}$$

In general, if Y as well as N is estimated (whether N is known or not) the (approximately unbiased) ***π-weighted estimator*** of $\bar{Y}$ may be written as:

$$\hat{\bar{Y}}_w = \frac{\hat{Y}_\pi}{\hat{N}_\pi} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i}} \tag{1.20}$$

where $\hat{N}_\pi = \sum_{i=1}^{n} \frac{1}{\pi_i}$ is the π-estimator of N (Särndal *et al.* 1992:182). This weighted estimator of $\bar{Y}$ is not unbiased, because its denominator is not fixed, i.e., it is a ratio estimator (Kish 1965:67).

1.3.1.3. The Finite Population Correction Factor (fpc)

For a simple random sample of size n selected without replacement from an *infinite* population, (or for a simple random sample selected *with replacement* from a *finite* population) it is well known that the variance of the mean is $\frac{\sigma^2}{n}$ where $\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}\left(y_j - \bar{Y}\right)^2$ (Cochran 1977:24). The only change in this result when the population is finite (when sampling without replacement) is the introduction of the factor $1-f = \frac{N-n}{N}$. This factor is usually called the *fpc* or the ***finite population correction***. Note that the sampling fraction $f = \frac{n}{N}$ tends to zero, i.e., $1-f$

approaches 1 when the population is much larger than the sample ($N >> n$). Hence, for an "infinite population" or when sampling with replacement, the fpc disappears from variance formulas.

1.3.2. Stratified Sampling

Stratified random sampling is one of the most widely used techniques in sample surveys.

In ***stratified sampling*** the population is divided into non-overlapping sub-populations called *strata*. A probability sample is selected independently in each stratum. To obtain the full benefits of stratification, the population sizes of the strata must be known. (Cochran 1977:89.)

In ***stratified random sampling*** a simple random sample is selected independently from each stratum (Cochran 1977:91).

In ***proportionate stratified sampling***, the sampling fraction is the same in each stratum, i.e., the sample size from a stratum is proportional to the population size in the stratum.

In ***disproportionate stratified sampling***, the sampling fractions vary among the strata.

Notation: The finite population $U = \{1,\ldots,j,\ldots,N\}$ is partitioned into L sub-populations, called *strata* and denoted as $U_1,\ldots,U_l,\ldots,U_L$. A probability sample s_l of size n_l is selected from U_l according to a design $p_l(\cdot)$ $(l = 1,\ldots,L)$. The number of elements in stratum l, called the size of stratum l, is denoted as N_l where $\sum_{l=1}^{L} N_l = N$. (Särndal *et al.* 1992:101.)

Notation: The population mean in stratum l is $\bar{Y}_l = \frac{1}{N_l}\sum_{j=1}^{N_l} y_j$ and the population variance in stratum l is $S_l^2 = \frac{1}{N_l - 1}\sum_{j=1}^{N_l}\left(y_j - \bar{Y}_l\right)^2$. The *srs wor* mean and variance in stratum l are $\bar{y}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} y_i$ and $s_l^2 = \frac{1}{n_l - 1}\sum_{i=1}^{n_l}\left(y_i - \bar{y}_l\right)^2$ respectively.

1.3.2.1. Stratified Random Sampling

Perhaps the most important type of stratified sampling is when *srs wor* is applied in all strata. This sampling strategy will be denoted *str*.

1.3.2.1.1. The str Estimator of the Population Total

The *str* estimator of the population total $Y = \sum_{l=1}^{L} N_l \bar{Y}_l$, is:

$$\hat{Y}_{str} = \sum_{l=1}^{L} N_l \bar{y}_l \tag{1.21}$$

The variance of the estimator $\hat{Y}_{str}$ is:

$$V\left(\hat{Y}_{str}\right) = \sum_{l=1}^{L} N_l^2 \left(1 - f_l\right) \frac{S_l^2}{n_l} \tag{1.22}$$

where $f_l = \dfrac{n_l}{N_l}$ is the sampling fraction in stratum l.

An unbiased variance estimator is:

$$v\left(\hat{Y}_{str}\right) = \sum_{l=1}^{L} N_l^2 \left(1 - f_l\right) \frac{s_l^2}{n_l} \tag{1.23}$$

1.3.2.1.2. The str Estimator of the Population Mean

The *str* estimator of the population mean $\bar{Y}$ is:

$$\hat{\bar{Y}}_{str} = \frac{1}{N} \sum_{l=1}^{L} N_l \bar{y}_l = \sum_{l=1}^{L} W_l \bar{y}_l \tag{1.24}$$

where $W_l = \dfrac{N_l}{N}$ is called the *weight* of stratum l. In the case of stratification with *proportional allocation*, the n_l are chosen so that $W_l = \dfrac{N_l}{N} = \dfrac{n_l}{n} \equiv w_l$. Stratified sampling with proportional allocation yields a *self-weighting* sample (see section 1.3.3.3).

The variance of the estimator $\hat{\bar{Y}}_{str}$ is:

$$V\left(\hat{\bar{Y}}_{str}\right) = \sum_{l=1}^{L} W_l^2 \left(1 - f_l\right) \frac{S_l^2}{n_l} \tag{1.25}$$

An unbiased variance estimator is:

$$v\left(\hat{\bar{Y}}_{str}\right) = \sum_{l=1}^{L} W_l^2 \left(1 - f_l\right) \frac{s_l^2}{n_l} \tag{1.26}$$

In the case of proportional allocation, the variance (1.25) reduces to (Cochran 1977:93):

$$V\left(\hat{\bar{Y}}_{str}\right)=(1-f)\sum_{l=1}^{L}W_l\frac{S_l^2}{n} \tag{1.27}$$

One advantage of stratified sampling is that stratification *may* produce a gain in *precision* (Kish 1965:76). In proportionate stratified sampling, precision may be increased to the degree that homogeneous strata (with respect to the survey variable) can be formed or alternatively, to the degree that strata can be formed with great heterogeneity among their means. Homogeneity within strata and heterogeneity among strata is affected by the choice of stratification variable: there is no gain in precision if the stratification variable is unrelated to the survey variable.

The ***stratification variable*** is the characteristic used for sub-dividing the population into strata.

The choice of stratification variable is limited to those variables for which the population distribution is known (Moser & Kalton 1971:91). If this requirement is not met, prior stratification is impossible. In such cases, the technique of *post-stratification* may be used.

1.3.2.2. Post-stratification

Post-stratification is an important technique that may produce significant gains in efficiency (Särndal *et al.* 1992:265). It is used when the sub-population *group* or "stratum" to which an element belongs is not known beforehand for the N population elements (Cochran 1977:134). However, from external sources, e.g., a previous census, accurate information about the population group *sizes* may be known. Suppose a probability sample is selected directly from the (unstratified) population and group membership is established for sampled elements only after the sampling has taken place. The groups are then called ***post-strata***.

<u>*Notation*</u>: Denote the sub-population groups (post-strata) by $U_1,\ldots,U_g,\ldots,U_G$ and let N_g be the size of U_g. If s is the sample drawn from U with the given sampling design, there is a corresponding partitioning of s into subsets $s_1,\ldots,s_g,\ldots,s_G$ where s_g is the part of s that falls in U_g and n_g is the size of s_g. The group size n_g is random[5] but the total sample size n is fixed.

[5] In the conditional inferences of subsequent chapters, these sub-group sizes will be treated as fixed.

<u>Notation</u>: The population mean in group g is $\bar{Y}_g = \frac{1}{N_g}\sum_{j=1}^{N_g} y_j$ and the population variance in group g is $S_g^2 = \frac{1}{N_g - 1}\sum_{j=1}^{N_g}(y_j - \bar{Y}_g)^2$. The *srs wor* sample mean and variance in population group g are $\bar{y}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} y_i$ and $s_g^2 = \frac{1}{n_g - 1}\sum_{i=1}^{n_g}(y_i - \bar{y}_g)^2$ respectively.

The ***post-stratified estimator*** of the population mean is:

$$\hat{\bar{Y}}_{pst} = \frac{1}{N}\sum_{g=1}^{G} N_g \bar{y}_g \tag{1.28}$$

There is disagreement in literature about the variance of $\hat{\bar{Y}}_{pst}$ and, in particular, the sampling distribution to which it should be related. There are two contenders (Holt & Smith 1979:34):

1. The distribution *conditional* on the stratum sizes n_g actually attained in the sample under study
2. The *unconditional* distribution determined by all possible samples of fixed size n

The conditional variance of $\hat{\bar{Y}}_{pst}$ is (Holt & Smith 1979:34):

$$V\left(\hat{\bar{Y}}_{pst} \mid n_g\right) = \sum_{g=1}^{G} W_g^2\left(1 - f_g\right)\frac{S_g^2}{n_g} \tag{1.29}$$

which is the usual variance for stratified samples (see (1.25)). The unconditional variance is obtained by averaging (1.29) over all possible values of n_g. This gives the approximate variance (Holt & Smith 1979:34):

$$AV(\hat{\bar{Y}}_{pst}) = \frac{1-f}{n}\sum_{g=1}^{G}\frac{N_g}{N}S_g^2 + \frac{1}{n^2}\sum_{g=1}^{G}(1 - \frac{N_g}{N})S_g^2 \tag{1.30}$$

assuming that $n_g \neq 0$ for all g. (If $n_g = 0$ for some g then neither variance can be employed directly. One practical solution is to pool or collapse similar groups.) For a variance estimator, the S_g^2 may be replaced by s_g^2.

The first term on the right-hand side of (1.30) is equal to the variance of a proportionate stratified random sample in (1.27) while the second term becomes negligible for large n_g. Hence, the (unconditional) variance of the *pst*-estimator may approach that of a proportionate stratified random sample of the same size (provided that n_g is reasonably large) but it cannot be less (Kish 1965:90).

Authors who support the use of the unconditional variance include Hansen, Hurwitz and Madow (1953), Raj (1968), Cochran (1977) and Kish (1965), while supporters of the conditional variance include Durbin (1969), Cox and Hinkley (1974), Kalton in his discussion of Smith (1976), Royall (1971) and Royall and Eberhardt (1975). Holt and Smith (1979:35) compare the two forms of inference both empirically and theoretically and "come down firmly in favour of conditional inferences". They come to the conclusion that when comparing sampling strategies before the sample is drawn, the unconditional variance should be used, whereas for inferences after the sample has been drawn the conditional variance is appropriate.

Throughout this dissertation, bias and variance expressions will be given *conditional* on the realised sample, i.e., conditional inferences will be employed.

1.3.3. Other Sampling Strategies

1.3.3.1. Systematic Sampling

Systematic sampling schemes offer several practical advantages, particularly simplicity of execution and ease of administration.

> ***Systematic selection*** denotes the selection of sampling units in sequences separated on lists by the interval of selection (Kish 1965:21).

Systematic sampling will not be used in this dissertation.

1.3.3.2. Cluster Sampling

In many sample surveys, direct element sampling is not possible or not desirable because of various reasons. For example, a suitable sampling frame that identifies each and every population element may not be available or can only be obtained at great cost. Other reasons have to do with the geographical location of population elements: if they are scattered over a wide area, direct element sampling may result in a widely scattered sample which, in face-to-face interview surveys, results in high travel expenses. Administrative efficiency of the survey is also affected: efficient supervision of the fieldwork in such a sample may be difficult which may lead to high non-response rates and severe measurement errors. (Särndal *et al.* 1992:124.)

In ***cluster sampling*** the finite population is grouped into disjoint sub-populations, called clusters. A probability sample of clusters is selected and every population element in the selected clusters is surveyed. (Särndal 1992:124.)

Cluster sampling involves a single stage of selection and is therefore also called *single-stage cluster sampling.*

1.3.3.3. Multi-stage Sampling

In some sample surveys, a single stage is required to select the sample (*single stage sampling*) but in most sample surveys it is desirable or essential to carry out the sampling in two or more stages (*multi-stage sampling*) (Moser & Kalton 1971:106).

In ***multi-stage sampling***, each stage has its own type of sampling unit: the first-stage sampling units are called ***primary sampling units*** (PSU's); the second-stage sampling units within the PSU's are called ***secondary sampling units*** (SSU's); the third-stage sampling units within the SSU's are called ***tertiary sampling units*** (TSU's), and so on. The sampling units in the last stage of sampling are called ***ultimate sampling units*** (USU's). (Särndal *et al.* 1992:125.)

In ***two-stage cluster sampling***, a probability sample of PSU's is drawn. From each selected PSU, a probability sample of SSU's (which are also USU's) is drawn. Each SSU is a cluster of elements and every element in the selected SSU's is surveyed. In ***two-stage element sampling***, every SSU is an element.

In *multi-stage element sampling* the USU's are elements.

In *multi-stage cluster sampling*, the USU's are clusters of elements and every element in the selected USU's is surveyed.

Complex sampling: In practice, samples are often selected in various stages using a combination of stratification, cluster sampling and simple random sampling. This is called *multi-stage stratified cluster sampling*, also known as complex sampling. (Stoker 1988:8.)

Notation: In stage I of multi-stage sampling[6], the population of elements $U = \{1,\ldots,j,\ldots,N\}$ is partitioned into N_I PSU's, denoted $U_1,\ldots,U_k,\ldots,U_{N_I}$. The index I will be used to identify entries associated with the PSU's. The number of population elements in PSU k is denoted N_k where $\sum_{U_I} N_k = N$. A sample s_I of PSU's is drawn from U_I. The number of PSU's in s_I is denoted as n_I.

[6] Detailed notation for subsequent stages of sampling will not be necessary for the purposes of this dissertation.

<u>*Notation*</u>: In stage II of two-stage element sampling, for every $k \in s_I$, a sample s_k of elements is drawn from U_k. The number of elements in s_k is denoted n_k, where $k = 1,2,\ldots,N_I$.

In multi-stage sampling, the calculation of estimates may be rather complex. By suitable choice of sampling fractions, it is often possible to keep the overall inclusion probabilities constant for different parts of the population, yielding a *self-weighting* design. The main advantage of a self-weighting design is that the sample data may be used unweighted (i.e., as observed) when calculating estimates and their variances, which leads to considerable simplification of the computations (Stoker 1988:16).

A single-stage sample may be called a ***self-weighting sample*** if each population element $j \in U$ has an equal inclusion probability, i.e., $\pi_j = \pi$ for all $j \in U$. A multi-stage sample may be called a self-weighting sample if each USU has an equal *total* inclusion probability (Stoker 1988:38).

Multi-stage sampling should be distinguished from *multi-phase sampling* (see section 1.3.3.6).

1.3.3.4. Unequal Probability of Selection Methods

The advantage of using epsem sampling is that it leads to *self-weighting* samples which lead to simplified estimators. However, unequal inclusion probabilities are often used in practice and for various reasons (Warwick & Lininger 1975:103):

1. To have sufficient sample sizes in small domains or to ensure adequate representation of "scarce" population elements
2. To control for size differences in PSU's by selecting PSU's with probability proportional to size (PPS)
3. To reduce the costs of sampling and/or interviewing by using a smaller sampling fraction in more expensive or less accessible segments of the population
4. To increase precision by using larger sampling fractions in strata with large variance

Unequal probability sampling may be employed in at least three ways (Warwick & Lininger 1975:105):

1. By using disproportionate allocation in stratified sampling

2. By deliberately assigning larger or smaller inclusion probabilities to special types of sampling units on the frame (without forming separate strata)

3. By assigning probabilities proportional to the size of sampling units (*PPS sampling*)

Two basic techniques may be used to compensate for unequal inclusion probabilities:

1. Assign a *weight* ω to each sample element where ω is proportional to the inverse of the (overall) inclusion probability of the sample element.

2. In multi-stage sampling, assign *compensating inclusion probabilities* in the various stages of the design, i.e., allow inclusion probabilities to vary within each stage but use a chain of inclusion probabilities which ultimately produces the same inclusion probability for each sample element. In this way, a self-weighting sample is produced (Kish 1965:21).

Four *implications* of the use of weights ω to compensate for unequal inclusion probabilities are (Stoker 1988:40):

1. It is possible to calculate unbiased estimators of population values such as the population total and population mean (if N is known)

2. The use of highly variable weights leads to an increase in the variance of estimators of population values

3. Frequency tables become meaningless

4. The application of statistical analysis techniques originally designed for unweighted data, for example CHAID (see section 3.2.2), may lead to incorrect results

1.3.3.5. Selection with Probability Proportional to Size Measures

In cluster sampling or multi-stage sampling, the total sample size n may be subject to unduly large variation if it is based on a random selection of clusters or PSU's that differ greatly in size (Kish 1965:217). For example, if the PSU's are chosen with equal probability $\frac{n_I}{N_I}$ and the same sampling fraction $\frac{n_k}{N_k}$ is applied within each selected PSU, an epsem sample is obtained ($\pi_i = \frac{n_I}{N_I} \times \frac{n_k}{N_k} = constant$). However, the ultimate sample size depends on which PSU's are selected in the first stage.

One method of obtaining greater control over sample size and which yields a self-weighting design, is to stratify the PSU's by size and select a sample in each group-size, probably with variable sampling fractions. An alternative procedure which also leads to a self-weighting design but gives complete control over sample size, is to select n_I PSU's with ***probability proportional to size (PPS)*** and then take a set number of elements n_k from each selected PSU. (Moser & Kalton 1971:111.)

A practical limitation of PPS sampling is that the PSU sizes must be known. If accurate and up-to date estimates are available these may be used; if not, it is better to use rough size measures M and PPS sampling than selecting PSU's with equal probabilities (Moser & Kalton 1971:112). The inclusion probabilities in sampling with probability proportional to size measures M are:

$$\pi_j = \frac{n_I M_k}{M} \times \frac{n_k}{M_k} = \frac{n_I n_k}{M} \tag{1.31}$$

where M_k is the estimated size measure of the k-th PSU and M is the estimated population size. Typical size measures are total assets or number of employees for a population of business firms, total acreage for a population of farms and total number of beds for a population of hospitals.

1.3.3.6. Multi-phase Sampling

In some surveys there is little or no prior knowledge about the population – knowledge that may be used to improve precision – so that only extremely simple designs, for example simple random sampling can be used. An option in such cases is to use the technique of *two-phase sampling* (also called *double sampling*). The aim is the creation of a highly informative frame, not for the whole population (this may be too expensive) but for a part of the population. A sub-sample is then drawn from the first phase-sample using the information collected in the first phase.

Multi-phase sampling refers to the sub-selection of the final sample from a pre-selected larger sample that provides information for improving the final selection (Kish 1965:21).

If π_i^* denotes the inclusion probability for sample element i under two-phase sampling (Särndal *et al.* 1992:347) the π^*-estimator of the population total is given by:

$$\hat{Y}_{\pi^*} = \sum_{i=1}^{n} \frac{y_i}{\pi_i^*} \qquad (1.32)$$

The theory of two-phase sampling is used in the treatment of non-response (see Chapter 4).

1.4. SYSTEMS OF INFERENCE

1.4.1. Model-based vs. Design-based Inferences

There are two currently competing approaches to statistical inference in surveys, namely the *classical* or *design-based* approach (also called the *randomisation* approach) and the *model-based* approach (Särndal 1978:27). The design-based approach is widely accepted in survey sampling *practice*, but in recent years some researchers, mainly those concerned with the *theoretical* basis of statistical sampling theory have challenged the use of design-based inferences and suggested a model-based alternative.

One important difference between the two approaches is the population to which inferences are made: design-based inferences are made to the finite target population, whereas model-based inferences are made to an infinite *super-population* (Kalton 1983b:178). Another important difference between design-based and model-based inferences lies in the element of randomness they utilise in order to give stochastic structure to inferences (Särndal 1978:27). Design-based inferences assume the *fixed population approach*: with each population element is associated a fixed but unknown real number y_j, the value of the survey variable y. In the fixed population approach, the randomisation introduced by the probability sampling mechanism provides the basis for statistical inference. (Hansen, Madow & Tepping 1983:776.) On the other hand, model-based inferences assume the *super-population* approach: with each population element is associated a *random variable* Y_j with a specified random structure. The N-dimensional distribution of $Y = (Y_1, \ldots, Y_j, \ldots, Y_N)$ is called the super-population model and is denoted as ξ. The actual value associated with a population element is treated as a realisation of the random variable Y_j. Usually ξ is indexed by an unknown vector of parameters, $\underline{\theta}$, the estimation of which is required as a prelude to making inferences about the finite population itself. (Särndal 1978:32.)

In the model-based approach, the super-population model ξ provides the basis for statistical inferences and probability sampling is therefore no longer necessary. However, proponents of model-based inferences advocate probability sampling as a safeguard against selection bias and to ensure "balanced" representation of the population, but sampling probabilities play no role in statistical inferences. The validity of statistical inferences now depends on the correct choice of model. (Smith 1976:192.)

Model-based inferences may have substantial advantages if the model is appropriate. For example, for certain models useful inferences can be based on quite small samples or skew (non-normal) distributions. However, Hansen, Madow and Tepping (1983:778) warn:

> *... if the assumed model does not accurately represent the state of nature, estimates of population values may be substantially biased, and statements about the sampling errors of those estimates may be very misleading.*

Kalton (1983b:186) concludes:

> *Models provide valuable insights that help to guide the choice of sample design, but most practitioners are reluctant to rely on them completely in either design or analysis. In most situations they prefer the robust inferences that come from the design-based approach. The main theoretical basis for this resides in the large samples typical of most surveys; in addition the multipurpose nature of surveys and the need for timely estimates present major practical obstacles to the widespread use of model-based estimators in survey research. Model-based estimators are, however, employed in some situations and their use seems almost certain to increase in future ...*

1.4.2. Model-assisted Design-based Inferences

The most attractive aspect of the design-based approach to statistical inference is that unverifiable assumptions about the distribution of characteristics in the finite target population are unnecessary, provided, of course, that the sample size is large (Rubin 1983:123). However, the drawback of the design-based approach is that when deviations from the probability sample occur in the form of non-sampling errors such as coverage error, measurement errors and non-response, pure design-based (model-free) inferences are no longer possible (see section 2.3.1). Specifically, in the case of non-response, the key ingredient of the design-based approach, namely a known probability distribution governing which values are observed and which are unobserved, is lost.

Two ways around this difficulty are available (Little & Rubin 1987:53):

1. Some modelling assumptions about the non-responding portion of the population are made, e.g., that the means of y in the responding and non-responding portions of the population are equal

2. No assumptions are made about the distribution of y in the population but a distribution is assumed for the *response mechanism* (see section 2.3.1)

Assumptions of the former kind relate to models for the y values and hence, to the model-based approach in surveys with complete response. The latter approach, which formulates a response distribution in addition to the sampling distribution, is a more direct extension of randomisation inference to the case of non-response. The term *quasi-randomisation inferences* was coined by Oh and Scheuren (1983) to describe this approach.

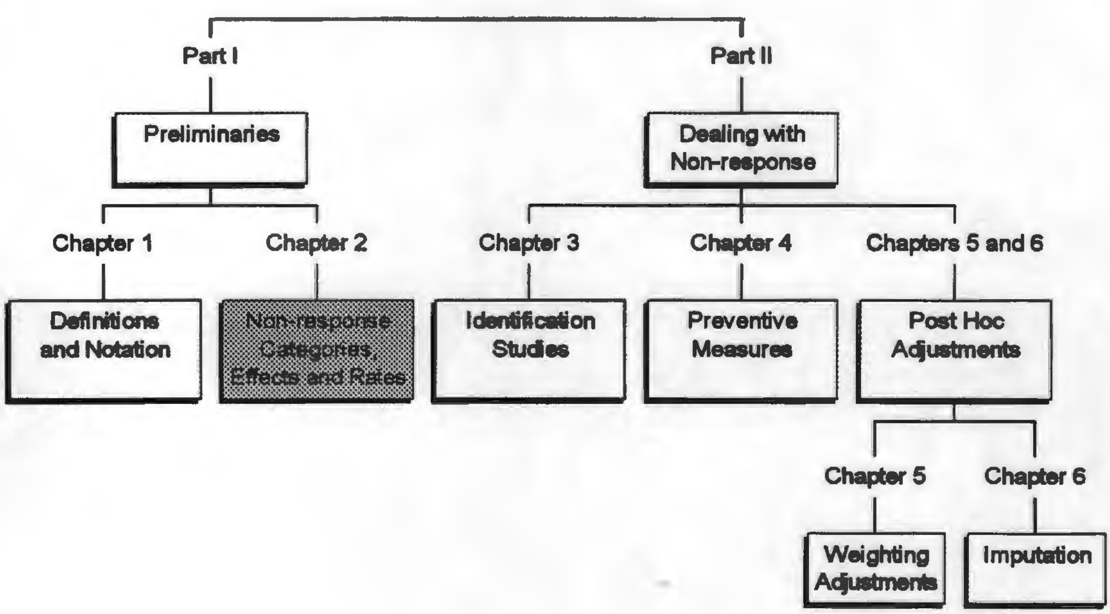
Model-assisted design-based inferences (or *quasi-randomisation inferences*) are based on the distribution obtained by combining the known model for the sampling mechanism with an assumed model for the response mechanism (Särndal *et al.* 1992:538).

The assumed distribution for the response mechanism typically involves assumptions that the data are missing across the entire sample or across specified sub-populations (see Chapters 5 and 6). Such assumptions can usually not be verified. Consequently, in the model-assisted design-based approach, the validity of inferences about the population depends on the truth of unverifiable assumptions. This is an undesirable situation which cannot be solved using a design-base: there is no ideal solution to the problem of non-response, except not to have any non-response in the first place!

The quasi-randomisation approach to inference will be followed in the remainder of this dissertation.

CHAPTER 2

NON-RESPONSE CATEGORIES, EFFECTS AND RATES



CHAPTER OUTLINE

2.1 INTRODUCTION

2.2 SOURCES OF UNIT NON-RESPONSE

2.3 THE EFFECTS OF UNIT NON-RESPONSE

2.4 CALCULATION OF RESPONSE RATES

2.5 INTERPRETATION OF RESPONSE RATES

CHAPTER 2

NON-RESPONSE CATEGORIES, EFFECTS AND RATES

The theory is based essentially on the textbook situation of "urns and black and white balls", and, while in agricultural and industrial sampling the practical situation corresponds closely to its theoretical model, the social scientist is less fortunate. He has to sample from an urn in which some of the balls properly belonging to it happen not to be present at the time of selection, while others obstinately refuse to be taken from it.

C.A. Moser & G. Kalton, Survey Methods in Social Investigation

2.1. INTRODUCTION

Non-response, a non-sampling error peculiar to probability samples[1], may "seriously impair the usefulness of a survey and undo the work that went into the sample" (Deming 1953:33). The uninitiated may wonder why non-response is a serious problem, since, with a non-response rate of say, 10%, a considerable portion of the original sample still remains. Experience shows, however, that the non-responding portion of the sample often differs from the rest - certainly, one should never just *assume* that it does not - and these differences are often related to the subject matter of the survey (Moser & Kalton 1971:166). In such cases the responding portion alone does not constitute a probability sample from the population so that estimators based only on respondent data must be assumed to be biased.

In section 2.2, the many diverse reasons why sample elements do not respond are grouped into a few meaningful categories. These categories of unit non-response are useful because (1) they have separate effects on survey results and (2) they require separate treatment. Firstly, a sample element who *refuses* an interview may be more unlike the responding portion of the sample (and have a more serious effect on survey results) than a sample element who is *temporarily away from home*. Secondly, improved operational procedures may be effective in dealing with *temporarily unavailables* but may have no effect on hard-core *refusals*.

In section 2.3, the effect of non-response on estimates of the population mean and total is examined. Firstly, it is shown that, in the presence of non-response,

[1] Non-response is not a problem in non-probability samples, since elements who do not wish to participate are simply replaced by those who are willing to participate.

unbiased design-based estimators are non-existent. Next, analytical expressions are derived which show how non-response bias produces its effects. These expressions are useful because they show that when tackling the non-response problem, a two-pronged approach is necessary: (1) reducing the non-response rate to the minimum possible level (within time and budget constraints) and (2) examining the differences between respondents and non-respondents in the population or in specified sub-groups of the population.

In section 2.4, the different types of "response rates" which can be found in literature and in research reports are examined with the purpose of establishing a standard definition of the response rate. It is shown that, what is calculated by many researchers as the response rate, is actually a type of operational "response" rate called a *completion rate.* In section 2.5, various issues surrounding the *interpretation* of response rates are addressed.

2.2. SOURCES OF UNIT NON-RESPONSE

Unit non-response occurs for a variety of reasons and in various degrees of "terminality". Often, a refuser on one occasion can be converted to a respondent (albeit a reluctant respondent) on another. If accurate records are kept of the outcomes of each attempted interview - e.g., by means of an accountability table as in Figure 2.4 - the sources of non-response in the survey can be identified and utilised to control and reduce non-response, to measure the non-response rate and to estimate its effects on survey results (Kish 1965:532).

2.2.1. Unit Non-response Categories

In Figure 2.1, possible categories of unit non-response are suggested for face-to-face interview surveys of households. Since the different methods of data collection generate different types of non-response, separate categories are suggested in Figures 2.2 and 2.3 for mail surveys and telephone surveys. The categories in Figures 2.1 to 2.3 may not all be applicable in a particular survey and other possibilities may exist, depending in each survey on the unique circumstances and characteristics giving rise to non-response. For example, terminology may have to be adjusted when other types of units such as business organisations, farms or hospitals are surveyed. In Figures 2.2

and 2.3, the sampling frame is assumed to consist of lists of named individuals and addresses or telephone numbers.

In many surveys one or more follow-up calls are made in an attempt to convert some non-respondents into respondents. In such surveys, the initial reason for non-response may differ from the final reason. For example: the initial reason for non-response may be *not-at-home* but the final reason may be *refusal.* All categories in Figures 2.1 to 2.3 refer to the final reason for non-response.

Figure 2.1 Categories of unit non-response in face-to-face interview surveys of households

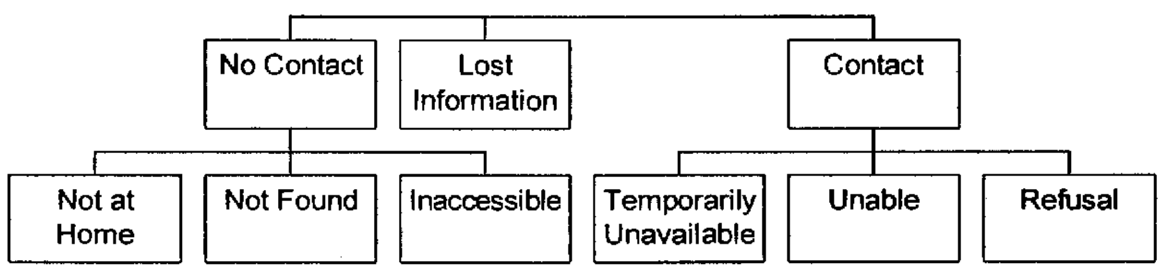


Figure 2.2 Categories of unit non-response in mail surveys of individuals

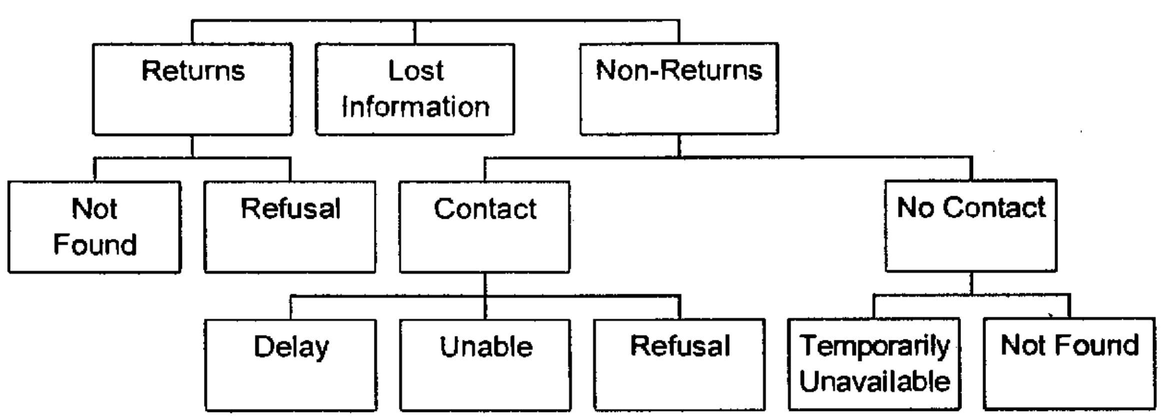
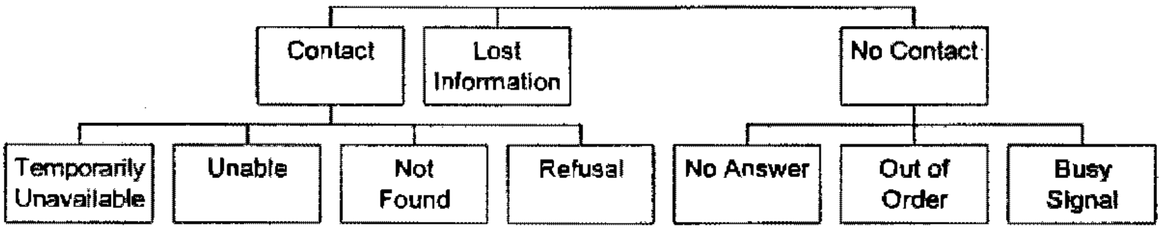


Figure 2.3 Categories of unit non-response in telephone surveys of individuals (directory sampling[2])



Figures 2.1 to 2.3 focus on sources of unit non-response, because reasons for item non-response are seldom reported (Lessler & Kalsbeek 1992:122). Some reasons for item non-response are: the respondent *refuses* to respond to some items, especially sensitive items such as age or income items; the interview is *terminated* after partial completion but some responses are still usable; the respondent does not have the information needed to respond to a particular item or the question is not clearly understood and therefore skipped; the interviewer or the respondent skips items because a branch point is missed[3]; the interviewer fails to ask a question or record a response and faked or unusable responses are detected during editing.

The application of the suggested classification system is subject to error. Sometimes the various categories are indistinguishable, for example, a sample element may not answer the doorbell as a way of *refusal* and may then be recorded as a *not-at-home* (Platek 1977:197). In mail surveys, the problem of indistinguishable categories is particularly prevalent since one cannot determine the reasons for the non-return of questionnaires, except perhaps in a face-to-face or telephone interview follow-up.

2.2.2. Eligibility of Sample Elements

It is important to note that all categories of non-response in Figures 2.1 to 2.3 refer to *eligible* elements only since, according to the definition of the response rate in section 2.4, *non-eligible* elements do not contribute to non-response in a survey. Whether or not an element is eligible for the survey is closely tied to its membership of

---

[2] In RDD methods the "not found" category among the contacted sample elements does not exist.

[3] See Messmer and Seymour (1982) for a discussion of the effects of branching on item non-response.

the target population[4], making a clear definition of the target population essential. For example, whether or not sample elements who are absent from home for the entire period of the survey (for example those on an extended holiday) should be treated as eligible, will depend on the definition of the population. Vacant housing units should be treated as non-eligible in surveys of *occupied housing units*, but in surveys where sampling is done from a list of named individuals and addresses, vacant units should contribute to non-response. The same holds for movers in mail surveys: if specific individuals were sampled but could not be found, they should be made to contribute to non-response.

In practice, the decision on the eligibility of sample elements is often made by an interviewer on the basis of observation, a brief screening interview or (after the interview has been completed) from more detailed information on the completed questionnaire (Madow, Nisselson & Olkin 1983:16). The difficulty is in determining whether those households or individuals that were not contacted are eligible for the survey. For example, if no one is at home after repeated calls in a face-to-face or telephone interview survey, the dwelling may or may not be vacant and hence, may or may not be eligible for the survey. In random digit dialling methods, some discontinued or unassigned numbers (non-eligible for the sample) may sound a normal ringing tone instead of giving a recorded message that the number has been discontinued or is not in use. Furthermore, some numbers where no answer is obtained may be for business or other units non-eligible for the survey.

2.2.3. Non-contacts

In some surveys, sample elements may not respond because they have not been contacted: they may be out at the time of each call, they may not have received the mailed questionnaire or their telephones may be out of order. In Figure 2.1, the *non-contact* category is understood to apply when there has been no contact with *any* member of the household. If a non-eligible member of a sample household is found at

[4] Ideally, as stated in section 1.2.9, there should be no difference between the target population and the survey population. If there is a difference, eligibility status of sample elements will depend on the definition of the survey population.

home, this member may provide information about the status of the eligible household member, for example, *temporarily unavailable*, *not at home* or *refuses* to respond.

The non-contact problem is primarily operationally oriented (Platek 1977:197) and is usually treated with operational solutions, such as allocating more time and resources to data collection, increasing the number of repeated calls that interviewers make on sample elements, improving interviewer training, reducing the interviewer burden or varying the time of making calls (see Chapter 4). The various reasons for non-contacts can be summarised in the three categories: eligible sample element *not at home*, *not found* or *inaccessible*. Although these three categories refer specifically to the reasons for non-contacts in face-to-face interview surveys (see Figure 2.1) they may also apply to some extent in mail and telephone surveys.

2.2.3.1. Not-at-home

If a face-to-face interview survey is conducted with a maximum of one call for each sample unit, the not-at-home category will form the majority of the non-response. Call-backs on sample units where no-one was found at home, will reduce the size of this category (see Chapter 4). The at-home patterns of the target population should be considered when calling or scheduling interviews: travelling salespeople, students and families where both parents work are harder to reach than housewives, farmers, families with young children and senior citizens. The timing of calls is important: calls made during the day are less likely to be successful than calls made over weekends and evenings. When scheduling face-to-face interview follow-ups, information may be obtained from neighbours or from non-eligible members of the household to determine a more appropriate time to re-call. Seasonal variation should also be considered when planning an interview, for example, the entire household may be absent for the length of the survey period so that call-backs are futile. (Kish 1965:532.)

Other factors affecting the number of not-at-homes in recent years (at least in European countries) have been smaller family sizes, greater mobility and a larger amount of spare time which is spent outdoors (Bethlehem & Kersten 1985:289).

The proportion of not-at-homes increases in surveys where the sample element is uniquely designated, for example the "head of the household" or the "oldest female".

These sample elements will be more difficult to find than when a range of possible respondents is allowed, such as "any person aged 18 or over".

In mail surveys, sample elements who are on holiday during the survey period or who are for any other reason *not at home*, may discard the questionnaire when they return, because it is "too late to complete". The advantages and disadvantages of setting a deadline date for the return of questionnaires have been debated to some extent (see section 4.2.11).

2.2.3.2. Not Found

In some surveys, sample elements cannot be found because the address could not be traced, the dwelling is vacant or does not exist or the sample element has moved or died[5]. The size of this category is affected by the quality of the frame in terms of accuracy and completeness and the amount of auxiliary information it contains (Lessler & Kalsbeek 1992:104). For example, the size of this category will be small if the frame contains complete and up-to-date addresses with accompanying maps as well as auxiliary information such as work address or telephone number or address of close friend or relative that may be used to trace individuals. In mail surveys, some of the non-respondents that belong to this category may be identified if the unopened envelopes are returned with or without a forwarding address. Unfortunately, if a mail questionnaire is not returned, the reason for non-response cannot be determined.

In face-to-face interview surveys and mail surveys, the necessity of tracing movers may be avoided by sampling *occupied dwellings* or *addresses* and not individuals (Lessler & Kalsbeek 1992:105) and in telephone surveys, random digit dialling instead of directory sampling may be used to avoid tracing.

2.2.3.3. Inaccessible

Some sample elements may be inaccessible to interviewers because of a "dangerous neighbourhood" or a "vicious watchdog" or because of their physical location, for example rural dwellers in mountains of Kwa Zulu-Natal or individuals in prisons. In some neighbourhoods and apartment blocks, access is security-controlled

---

[5] According to Groves (1989:139) deaths are most often treated as non-eligible cases, depending on the time of death. For example, if death occurred after sample selection, it should be classified as non-response.

(Lessler & Kalsbeek 1992:104). Bad weather during the survey period may also prevent interviews from taking place.

2.2.4. Contacts

2.2.4.1. Temporarily Unavailable

Sometimes contact is established at one of the calls but the sample element is temporarily unavailable for the interview. The sample element may be too busy, tired or ill at the time, but is willing to be interviewed at a later time (Kish 1965:533). The sample element has therefore not categorically *refused* the interview and may still be persuaded to respond on subsequent calls. On the other hand, an indefinite delay should be categorised as a refusal.

The size of this category may be reduced by improving interviewer persuasion techniques through efficient training, by using an advertising campaign to promote the aims of the survey and by establishing good relations between the survey organisation and the public (Platek 1977:198).

2.2.4.2. Unable

Some sample elements may be unable to respond because of physical, mental, emotional or language problems (Lessler & Kalsbeek 1992:125). For example, in mail surveys the illiterate are excluded automatically. If the *unable* category is large, the exact reason should be identified and attempts at obtaining responses should be made by using more suitable methods (Kish 1965:534). For example, where there is a language barrier a translator may be appointed by the survey organisation and where there is a sizeable portion of illiteracy among the population in the mail survey, responses should rather be obtained through face-to-face interviews. However, these issues should preferably be resolved before the survey is conducted, i.e., during the planning phase of the survey.

2.2.4.3. Refusal

There are various degrees of refusals, ranging from temporary (the *resisters* or *polite refusers*) to permanent (the *hard-core refusers*). The resisters are those who may be persuaded to respond by more intensive interviewing efforts, such as call-backs in face-to-face interview and telephone surveys or reminder letters in mail surveys. On the other hand, the hard-core refusers are those who cannot be persuaded to respond. No operational procedure will reach this category of non-response and its biasing effects will have to be incorporated into the survey results. (Of course, the non-contact category may also contain a number of refusers who will go undetected.) In mail surveys it is difficult to distinguish between temporary and permanent refusals.

In general, it is more difficult to determine the reasons for refusals than for non-contacts. In some surveys, sample elements are less amenable to respond because of some cultural, social or demographic characteristics which make the questions seem embarrassing, "personal" or irrelevant. Lengthy questionnaires and difficult questions may encourage refusals. Other factors which cause refusals are fear, distrust, fatigue, perceived invasion of privacy, disruption of leisure time, sensitivity about the subject, apathy and lack of time (Lessler & Kalsbeek 1992:105). Moser and Kalton (1971:174) believe that "common-sense" should be used to minimise refusals:

> *The steps a surveyor can take to minimise* refusals *are in the main matters of common sense. It will always help to keep the questionnaire as brief as possible so that the burden on the respondent is at a minimum, and to aid him in giving information, perhaps with the inducement of financial rewards. But much will depend on the sponsorship and purpose of the survey itself, on the interviewers, on the questions, and on the general approach.*

The ethics of offering financial rewards is debatable - and the issue of measurement error when offering financial rewards needs to be addressed (see section 4.2.10).

Since the achieved level of response depends very much on the interviewer and the way he/she presents the survey, it is important that interviewers are trained in public relations and have good communication and motivational skills. However, unwilling sample elements who are relentlessly being pursued by a persistent interviewer may give inaccurate information (leading to measurement errors) rather than continue to refuse. Dalenius (1961:2) warns that "participation enforced by means of compulsion *can* result in errors of reporting [measurement errors] which have consequences more serious than those of non-response".

2.2.5. Lost Information

Completed questionnaires may get lost or destroyed during or after data collection (Kish 1965:534). In mail surveys, some questionnaires may be unusable because of poor quality or illegible handwriting. In face-to-face interview or telephone surveys, some interviewers may have faked responses to reduce the size of non-response in their assignment. If this is detected in time, the elements should be re-interviewed or the real reason for non-response should be determined. In mail surveys, some questionnaires get lost in the mail before or after they are received and completed by the sample elements.

2.3. THE EFFECTS OF NON-RESPONSE ON SURVEY ESTIMATES

Non-response affects a survey in at least three ways[6]: (1) estimators are less accurate because of an increase in their *MSE*; (2) the cost of the survey is increased (unless the non-response is ignored completely) and (3) the univariate and multivariate distributions of survey variables are distorted so that the distributions of means, variances, covariances and other estimators are affected. The increase in the *MSE* is brought about by (a) the introduction of a bias component due to non-response and (b) an increase in the variance component due to the sample size being reduced from n to n_r where n_r denotes the number of respondents in the sample. The increase in survey costs arises from the additional data collection efforts which are needed to reduce the non-response rate (see Chapter 4) and from using various statistical methods to deal with non-response, such as weighting and imputation (see Chapters 5 and 6).

It is shown in this section that, when non-response is present in a survey, design-based inferences can no longer be used to obtain unbiased estimators. Thereafter, the bias resulting from non-response is examined in three situations: (1) confining the survey analyses to the respondent data, (2) attempting to increase the response rate to the highest level possible (subject to time and budget constraints) and (3) replacing non-respondent values by estimates obtained through an imputation or substitution procedure. Expressions for the non-response bias of estimators of the population mean and total are derived under two views of non-response, namely the

[6] The effects of item non-response will not be discussed separately from the effects of unit non-response, however, under the assumption that there is only one survey variable, unit and item non-response become equivalent.

deterministic and *stochastic* views. The cost-implications of non-response in surveys will not be discussed in this dissertation.

Some new terminology and notation which were not covered in Chapter 1, are introduced in section 2.3.1.

2.3.1. Lack of Unbiased Estimators

In the absence of non-response, the probability sampling design $p(s)$ describes how the sample s is generated from the finite population U. The inclusion probabilities π_j are known and are strictly positive for all $j \in U$. But in the presence of non-response, the selection of a sample s from the population U is followed by the "selection" of a *response set* from the sample.

The ***response set***, denoted s_r, is the sub-set of the sample s for which acceptable responses for y are obtained (Särndal *et al.* 1992:557). In this dissertation, the assumption is made that the response set is non-empty.

Notation: Given that the sample s was selected, let $\mathcal{R} = \{s_{r_1}, s_{r_2}, \ldots, s_{r_F}\}$ denote the set of all possible response sets s_r and let S_r denote the random variable taking set values $s_r \in \mathcal{R}$.

In the presence of non-response, the probability sampling design must be supplemented by a model which describes how the response set is generated, i.e., a *response mechanism* must be specified.

The ***response mechanism*** $p(s_r \mid s)$ specifies the probability that the response set s_r is generated, given that the sample s was selected.

In the case of a ***probability response mechanism***, each possible response set $s_r \in \mathcal{R}$ is assumed to have a known and non-zero conditional probability of being realised $P(S_r = s_r \mid s) = p(s_r \mid s) > 0$ or, equivalently, each element $j \in U$, if it were included in the sample s, is assumed to have a known and non-zero probability φ_j to respond.

If $p(s_r|s)$ were known for all $s \in \mathcal{L}$, properties of any estimator $\tilde{y}(S, S_r)$ could be calculated based on respondent data only, that is one would be able to calculate the expected value (see equation 1.3):

$$E[\tilde{y}(S,S_r)] = \sum_{s \in \mathcal{L}} \sum_{s_r \in \mathcal{R}} p(s)p(s_r \mid s)\tilde{y}(s,s_r) \tag{2.1}$$

and variance (see equation 1.4):

$$V[\tilde{y}(S,S_r)] = \sum_{s\in\mathcal{S}}\sum_{s_r\in\mathcal{R}} p(s)p(s_r|s)\left[\tilde{y}(s,s_r) - E\{\tilde{y}(S,S_r)\}\right]^2 \qquad (2.2)$$

The problem with non-response is that, except in truly exceptional cases, $p(s_r|s)$ is an unknown distribution and consequently, standard techniques for obtaining unbiased estimators do not work (Särndal *et al.* 1992:558). To make statistical inferences using respondent data, model-assumptions must be made about the distribution for $p(s_r|s)$, for example, that the elements respond independently of each other or that two or more similar elements have the same probability of responding (see Chapter 5).

The type of model used for the response mechanism will depend on whether the *deterministic* view of non-response (also called the "fixed response model") or the *stochastic* view of non-response (also called the "random response model") is assumed.

2.3.1.1. The Deterministic View of Non-response

According to the deterministic view of non-response, the population of N elements is assumed to consist of two sub-groups: a respondent sub-group and a non-respondent sub-group. The respondent sub-group consists of N_r elements that will always respond on conceptual replications of the survey under the same essential survey conditions. The non-respondent sub-group consists of N_{nr} elements that will never respond. The sub-group sizes N_r and N_{nr} are fixed. The response probabilities of the N_r elements in the respondent sub-group are $\varphi_j = 1$ for $j = 1,\ldots,N_r$ and the response probabilities for the N_{nr} elements in the non-respondent sub-group are $\varphi_j = 0$ for $j = 1,\ldots,N_{nr}$.

This view of the response mechanism certainly simplifies the analytical approach to non-response treatment (respondents and non-respondents are treated as separate study domains) but it is somewhat naive and "an oversimplification of reality" (Cochran 1977:360). Other factors, such as the subject of the survey, the method of data collection, interviewer workload, steps taken by the fieldworker to solicit response, timing of the interview and the frame of mind of the sample element play a part in determining whether or not a population element will respond (Lessler & Kalsbeek 1992:132).

2.3.1.2. The Stochastic View of Non-response

Under the stochastic view of non-response, whether or not an element responds, is assumed to be a random variable whose outcome is determined by various factors, such as those mentioned in section 2.3.1.1. Each population element $j \in U$, if selected for the sample, is assumed to have a response probability φ_j $(0 \le \varphi_j \le 1)$ which may differ among population elements. If a *probability response mechanism* is assumed, φ_j is assumed strictly positive and known for all elements in the population and estimation and analysis can proceed just as if *two-phase* sampling has taken place. In the case of a probability response mechanism, an unbiased estimator of the population total is given by the π^*-estimator (see equation 1.32):

$$\hat{Y}_{\pi^*} = \sum_{i=1}^{n_r} \frac{y_{r_i}}{\pi_i^* \varphi_i} \tag{2.3}$$

The problem of estimating the response probabilities φ_i in (2.3) is a substantial one. Often, the data are assumed to be missing at random within *specified sub-groups* of the population defined on the basis of auxiliary information available for the entire population (Kalton 1983b:182). In this approach, elements are classified into sub-groups and assumptions are made regarding response probabilities within the sub-groups. The estimated response probabilities are then used to adjust for non-response. Non-response adjustments based on assumed response probabilities within sub-groups will be discussed in Chapter 5.

2.3.1.3. Differences Between the Deterministic and Stochastic Views of Non-response

Table 2.1 summarises various differences between the deterministic and stochastic views of non-response.

Table 2.1 Differences between the deterministic and stochastic views of non-response

Deterministic View	Stochastic View
Constant φ_j over replications of the survey	Random φ_j (φ_j may vary over replications of the survey)
$\varphi_j = 0$ or $\varphi_j = 1$	$0 \le \varphi_j \le 1$
Fixed N_r and N_{nr}	N_r and N_{nr} may vary over replications of the survey
Non-probabilistic: Inferences apply only to the respondent sub-group	Assuming a known *probability* response mechanism ($0 < \varphi_j \le 1$), classical design-based inferences to the entire population are possible

Since the formulas for non-response effects are simplified analytically by assuming the deterministic viewpoint, the deterministic viewpoint is an attractive approach to the non-response problem. However, while a probability response mechanism may be readily assumed under the *stochastic* view by letting $0 < \varphi_j \le 1$, under the *deterministic* view, the requirements of a probability response mechanism are not met ($\varphi_j = 0$ or $\varphi_j = 1$). Therefore, inferences made under the deterministic viewpoint apply only to the respondent stratum. Two ways around this problem are:

1. Assume that the population mean of the respondents is equal to the population mean of the non-respondents, i.e., assume that $\overline{Y}_r = \overline{Y}_{nr}$. This is a model-based approach and will not be used in this dissertation.

2. Consider the deterministic view as a *conditional* form of the stochastic view (given the actual response outcome, "response" or "non-response" for each population element) (Lessler & Kalsbeek 1992:162). Under this view, each population element is still allowed to have a random response probability φ_j. Estimation of the φ_j is possible by assuming, for example, that elements in the *population* (or in specified *sub-groups* of the population) respond with the same probability, which can be estimated by the sample response rate (within sub-groups) (Anderson 1979:108). This implies that φ_j is considered to be the *proportion* of population elements that would respond if they were selected for a particular survey, i.e., the population response rate R (see section 2.4).

The second option, considering the deterministic view as a conditional form of the stochastic view, will be employed throughout this dissertation. The effect of non-response bias under the stochastic viewpoint will be considered briefly and only in the present chapter.

2.3.2. The Bias of Non-response: Deterministic View

2.3.2.1. Notation and Definitions

Suppose the relative size of the respondent group is denoted as $R = \dfrac{N_r}{N}$ and the relative size of the non-respondent group is denoted as $\tilde{R} = 1 - R = \dfrac{N_{nr}}{N}$. R and $\tilde{R}$ can be called the *population unit response* and *non-response rate* respectively as defined in section 2.4. For simplicity, assume that an *srs wor* of size n is selected resulting in n_r responses from the respondent group but no data from the n_{nr} non-respondents. Define the population mean in the respondent sub-group:

$$\bar{Y}_r = \frac{1}{N_r}\sum_{j=1}^{N_r} y_{rj} \tag{2.4}$$

and the population mean in the non-respondent sub-group:

$$\bar{Y}_{nr} = \frac{1}{N_{nr}}\sum_{j=1}^{N_{nr}} y_{nrj} \tag{2.5}$$

The overall population mean can be written as:

$$\bar{Y} = R\bar{Y}_r + \tilde{R}\bar{Y}_{nr} \tag{2.6}$$

Define the respondent sample mean:

$$\bar{y}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} y_{ri} \tag{2.7}$$

and the non-respondent sample mean:

$$\bar{y}_{nr} = \frac{1}{n_{nr}}\sum_{i=1}^{n_{nr}} y_{nri} \tag{2.8}$$

The population variance in the respondent sub-group is defined as:

$$S_r^{\ 2} = \frac{1}{N_r - 1}\sum_{j=1}^{N_r}\left(y_{r_j} - \bar{Y}_r\right)^2 \tag{2.9}$$

and the population variance in the non-respondent sub-group as:

$$S_{nr}^{\ 2} = \frac{1}{N_{nr} - 1}\sum_{j=1}^{N_{nr}}\left(y_{nr_j} - \bar{Y}_{nr}\right)^2 \tag{2.10}$$

The sample variance in the respondent sub-group is defined as:

$$s_r^{\ 2} = \frac{1}{n_r - 1}\sum_{i=1}^{n_r}\left(y_{r_i} - \bar{y}_r\right)^2 \tag{2.11}$$

and the sample variance in the non-respondent sub-group as:

$$s_{nr}^{\ 2} = \frac{1}{n_{nr} - 1}\sum_{i=1}^{n_{nr}}\left(y_{nr_i} - \bar{y}_{nr}\right)^2 \tag{2.12}$$

In the presence of non-response, a researcher may (1) choose to do nothing or (2) use a statistical method to deal with non-response, such as estimation of the non-response bias and incorporating it into the results (see Chapter 3), substitution (Chapter 4), improved data collection procedures with the aim to increase the response rate or reduce the non-response rate (Chapter 4), weighting (Chapter 5) or imputation (Chapter 6).

In this section, expressions for the bias of non-response are considered in the following situations:

1. Do nothing
2. Aim to reduce the sample non-response rate to a minimum
3. Obtain estimates $\hat{\bar{Y}}_{nr}$ through some method such as imputation or substitution

2.3.2.2. Method 1: Do Nothing

2.3.2.2.1.Bias of the srs wor Estimator of the Population Mean

If no compensation is made for non-response[7], the *srs wor* estimator (see equation 1.16) of the population mean is the respondent sample mean $\bar{y}_r$. Its bias is given by:

[7] Note that ignoring missing data is equivalent to imputing the respondent mean for each non-respondent.

$$\begin{aligned} Bias(\bar{y}_r \mid n_r) &= E(\bar{y}_r \mid n_r) - \bar{Y} \\ &= \bar{Y}_r - \bar{Y} \\ &= \bar{Y}_r - (R\bar{Y}_r + \tilde{R}\bar{Y}_{nr}) \\ &= \tilde{R}(\bar{Y}_r - \bar{Y}_{nr}) \end{aligned} \tag{2.13}$$

Since, $E(\bar{y}_r \mid n_r) = \bar{Y}_r$, the respondent mean is an unbiased estimator of $\bar{Y}_r$, but is a biased estimator of $\bar{Y}$, that is:

$$E(\bar{y}_r \mid n_r) \neq \bar{Y}$$

unless $\bar{Y} = \bar{Y}_r (= \bar{Y}_{nr})$ in which case non-response is said to be "ignorable" (Rubin 1987:22).

The expression (2.13) shows that the non-response bias of the respondent mean is equal to the product of the population non-response rate and the difference between the population means for respondents and non-respondents.

An alternative expression for the non-response bias of $\bar{y}_r$ may be obtained when the respondent/non-respondent means vary across the categories of non-response (see section 2.2). Suppose the population can be divided into four groups, namely a respondent group, a non-contact group (NC), a not-available group (NA) and a refusal group (RF). Suppose the population means in these groups are denoted as $\bar{Y}_r$, $\bar{Y}_{NC}$, $\bar{Y}_{NA}$ and $\bar{Y}_{RF}$ respectively and the population sizes in these groups are denoted as N_r, N_{NC}, N_{NA} and N_{RF} respectively. If $N_{nr} = N_{NC} + N_{NA} + N_{RF}$, then the bias of non-response may be expressed as:

$$Bias(\bar{y}_r \mid \mathbf{n}) = \frac{N_{NC}}{N}(\bar{Y}_r - \bar{Y}_{NC}) + \frac{N_{NA}}{N}(\bar{Y}_r - \bar{Y}_{NA}) + \frac{N_{RF}}{N}(\bar{Y}_r - \bar{Y}_{RF}) \tag{2.14}$$

where $\mathbf{n}$ is the vector of realised sample sizes from each group.

According to Groves (1989:134):

> *The more complex expression is more enlightening because there is little prior belief that the values of the statistics for these different kinds of non-respondents are similar.*

Madow, Nisselson and Olkin (1983:19) support Groves in stating:

> *At present commonly used statistical methods of dealing with non-response do not depend on type of non-response; such methods need to be more fully developed in the future...*

Of course the expression (2.14) is very difficult to estimate in practice, since there is *usually* no way to obtain an estimate of the quantity $\overline{Y}_{RF}$.

2.3.2.2.2. Bias of the srs wor Estimator of the Population Total

In the absence of non-response, the *srs wor* estimator of the population total $Y = N\overline{Y}$ may be written as (see equation 1.12):

$$\hat{Y}_{srs} = N\bar{y} = N\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\sum_{i=1}^{n} y_i}{f} \tag{2.15}$$

In the presence of non-response, the application of the sampling fraction *f* to the sample total for the respondents gives an estimator:

$$\hat{Y}_1 = N\frac{1}{n}\sum_{i=1}^{n_r} y_{r_i} = \frac{n_r \bar{y}_r}{f} \tag{2.16}$$

with expectation:

$$E(\hat{Y}_1 \mid n_r) = \frac{1}{f}E\left[n_r \bar{y}_r \mid n_r\right] = \frac{N}{n}(n_r \overline{Y}_r) \tag{2.17}$$

If $\frac{n_r}{n}N \approx N_r$, the approximate bias of the estimator $\hat{Y}_1$ is:

$$\begin{aligned} Bias(\hat{Y}_1 \mid n_r) &= E(\hat{Y}_1 \mid n_r) - N\overline{Y} \\ &\approx N_r \overline{Y}_r - N\overline{Y} \\ &= -N_{nr}\overline{Y}_{nr} \end{aligned} \tag{2.18}$$

On the other hand, the estimator of the population total:

$$\hat{Y}_2 = N\bar{y}_r \tag{2.19}$$

has bias:

$$\begin{aligned} Bias(N\bar{y}_r \mid n_r) &= E(N\bar{y}_r \mid n_r) - N\overline{Y} \\ &= N\overline{Y}_r - N(R\overline{Y}_r + \tilde{R}\overline{Y}_{nr}) \\ &= N\tilde{R}(\overline{Y}_r - \overline{Y}_{nr}) \end{aligned} \tag{2.20}$$

which is comparable to $Bias(\bar{y}_r \mid n_r)$ in expression (2.13).

2.3.2.2.3. Effect of Non-response on the Estimation of the Population Variance

In the presence of non-response, the estimation of the population variance S^2 is usually based on the respondent sample variance, namely $s_r^{\,2}$ with expected value:

$$E(s_r^{\,2} \mid n_r) = S_r^{\,2} \qquad (2.21)$$

The approximate bias of $s_r^{\,2}$ as an estimator of S^2 is derived by Kalton (1983a:9) as:

$$Bias(s_r^{\,2} \mid n_r) \approx \tilde{R}(S_r^{\,2} - S_{nr}^{\,2}) - R\tilde{R}(\overline{Y}_r - \overline{Y}_{nr})^2 \qquad (2.22)$$

The first term on the right-hand side of (2.22) is negligible under the assumption that $S_r^{\,2} \approx S_{nr}^{\,2}$. According to Kalton (1983a:9) such an assumption is more realistic than the assumption that $\overline{Y}_r \approx \overline{Y}_{nr}$. The second term on the right-hand side of (2.22) reflects the effect of differences in the respondent and non-respondent means on the estimator of the population variance. If respondents and non-respondents have the same variance ($S_r^{\,2} = S_{nr}^{\,2}$), $s_r^{\,2}$ will underestimate S^2 ($s_r^{\,2}$ has a negative bias) unless $\overline{Y}_r = \overline{Y}_{nr}$, in which case the estimator $s_r^{\,2}$ has zero bias.

2.3.2.2.4. Interpretation of Non-response Bias of Estimators of the Population Mean and Total

The magnitude of the non-response bias of the estimators $\bar{y}_r$ of the population mean and $N\bar{y}_r$ of the population total can be examined by assuming various values for $\tilde{R}$ and $|\overline{Y}_r - \overline{Y}_{nr}|$. A number of conclusions can be made:

1. Non-response bias will be negligible when both (a) $\tilde{R}$ approaches zero and (b) $\overline{Y}_r \approx \overline{Y}_{nr}$. This means that if a small sample non-response rate coincides with small differences between respondents and non-respondents, bias due to non-response may be negligible.

2. Even when $\tilde{R}$ does not approach zero but where $|\overline{Y}_r - \overline{Y}_{nr}|$ approaches zero, bias due to non-response may be negligible. Conversely, even when $\tilde{R}$ is small, but $|\overline{Y}_r - \overline{Y}_{nr}|$ is large, non-response bias may be serious. This means that a large sample non-response rate but with small respondent/non-respondent differences

does not necessarily lead to serious bias. However, a large response rate does not guarantee small non-response bias if there are large respondent/non-respondent differences.

3. When a large $\tilde{R}$ coincides with large $|\bar{Y}_r - \bar{Y}_{nr}|$, estimators will be subject to substantial non-response bias.

Since the survey analyst usually has no direct empirical evidence on the magnitude of $|\bar{Y}_r - \bar{Y}_{nr}|$, the only situation in which he/she can have confidence that the bias is small, is when the non-response rate is low (Kalton 1983a:7). Efforts can be made to empirically estimate $|\bar{Y}_r - \bar{Y}_{nr}|$ in order to gauge the effect of non-response on estimates and correct for non-response bias (see Chapter 3), but these efforts may not always be successful.

2.3.2.3. Method 2: Reduce $\tilde{r}$ to a Minimum

One technique which is often used to attempt to reduce non-response bias is to increase the response rate to an "acceptable" level. However, from the discussion above, it should be clear that efforts to reduce the effect of non-response on survey estimators should not focus only on minimising the sample non-response rate, but that the *effect* of reducing the sample non-response rate on the size of respondent/non-respondent differences should also be considered. The sample non-response rate (formally defined in section 2.4) will be denoted as $\tilde{r}$.

According to Lessler and Kalsbeek (1992:142), respondent/non-respondent differences have rarely been considered in efforts to reduce non-response bias. Most researchers have focused only on reducing $\tilde{r}$ to a minimum. At first sight, it may seem that a reduction in $\tilde{r}$ will certainly lead to a smaller bias, but as $\bar{Y}_r$ and $\bar{Y}_{nr}$ can also change if $\tilde{R}$ changes, this is not necessarily so (Moser & Kalton 1971:169). Platek (1977:192) warns:

> *A reduction of nonresponse in the field does not necessarily ensure a reduction in bias. In fact, it can be shown that if the procedures for reduction of nonresponse are not well thought out and appropriately executed, the bias may not be reduced and could even be increased.*

It is not easy to predict in which way a reduction in $\tilde{r}$ will affect respondent/non-respondent differences. An increase in $|\bar{Y}_r - \bar{Y}_{nr}|$ may be expected as $\tilde{R}$ decreases when increased data collection efforts obtain responses from only one part of the survey population. On the other hand, a decrease in $|\bar{Y}_r - \bar{Y}_{nr}|$ may be expected as $\tilde{R}$ increases if non-respondents have similar values among themselves but differ collectively from the respondents. (Lessler & Kalsbeek 1992:142.) For example, suppose that all non-respondents in the population are either *hard-core refusers* or *temporary refusers* and through increased data collection efforts, all temporary refusers are persuaded to respond. It may be reasonable to expect larger $|\bar{Y}_r - \bar{Y}_{nr}|$, since the hard-core refusers are likely to differ more from respondents than the temporary refusers.

To examine the effect more closely, suppose that the non-respondent group in the population consists of two distinct groups: N_{RF} refusers and N_{NC} initial non-contacts with respective means $\bar{Y}_{RF}$ and $\bar{Y}_{NC}$. Suppose now $\tilde{R}$ is decreased through increased data collection efforts, but all those who are converted into respondents are non-contacts. The effect on non-response bias becomes clearer when considering the relative position of the means $\bar{Y}_{RF}$ and $\bar{Y}_{NC}$ with respect to $\bar{Y}_r$ (Moser & Kalton 1971:169). If $\bar{Y}_{RF}$ and $\bar{Y}_{NC}$ lie on opposite sides of the mean $\bar{Y}_r$, the bias due to non-response may actually be increased, while if $\bar{Y}_{RF}$ and $\bar{Y}_{NC}$ lie on the same side of the mean $\bar{Y}_r$, the bias may be reduced.

For example, suppose one is attempting to estimate the mean number of traffic fines received by university lecturers in the past year. Suppose the following values apply:

Respondents	Non-respondents	
	Refusers	Non-contacts
$\bar{Y}_r = 2{,}1$	$\bar{Y}_{RF} = 4{,}3$	$\bar{Y}_{NC} = 0{,}9$
$N_r = 800$	$N_{RF} = 100$	$N_{NC} = 100$

Here $\overline{Y}_{nr} = \frac{100(4,3)+100(0,9)}{200} = 2,6$ so that the bias of $\overline{y}_r$ as an estimator of $\overline{Y}$ is from (2.13):

$$Bias(\overline{y}_r \mid n_r) = 0,20(2,1-2,6) = -0,1$$

Suppose $\widetilde{R}$ is reduced from 20% to 10%, but suppose all of those who were converted into respondents are non-contacts, so that $\overline{Y}_r' = \frac{800(2,1)+100(0,9)}{900} = 1,97$ and $\overline{Y}_{nr} = \overline{Y}_{RF} = 4,3$. The bias of $\overline{y}_r'$ as an estimator of $\overline{Y}$ is from (2.13):

$$Bias(\overline{y}_r' \mid n_r) = 0,10(1,97-4,3) = -0,233$$

which is greater than the original bias of -0,1.

Although an increase in bias due to increased data collection efforts is possible, Moser and Kalton (1971:169) do state that such an increase in bias is rare. It is, however, important that the techniques used to reduce $\widetilde{r}$ are "well thought out" and "properly executed". (See the study by Wilcox (1977), discussed in section 3.6.9, for an empirical investigation of the interaction between refusal and non-contact bias which may lead to an increase instead of a decrease in non-response bias.)

2.3.2.4. Method 3: Imputation and Substitution

Assume for simplicity that a population census is undertaken. Suppose through some method (such as imputation or substitution) the N_{nr} missing values of y are estimated by $\hat{y}_{nrj} = z_j$ for $j = 1,2,\ldots,N_{nr}$. Suppose the mean $\overline{Z}_{nr}$ is used as an estimator of $\overline{Y}_{nr}$, that is:

$$\hat{\overline{Y}}_{nr} = \overline{Z}_{nr} = \frac{1}{N_{nr}} \sum_{j=1}^{N_{nr}} z_j \tag{2.23}$$

An estimator of $\overline{Y}$ may be defined as (Madow, Nisselson & Olkin 1983:25):

$$\hat{\overline{Y}} = R\overline{Y}_r + \widetilde{R}\overline{Z}_{nr} \tag{2.24}$$

with bias:

$$\begin{aligned} Bias(\hat{\bar{Y}} \mid N_r) &= E(\hat{\bar{Y}} \mid N_r) - \bar{Y} \\ &= (R\bar{Y}_r + \tilde{R}\bar{Z}_{nr}) - (R\bar{Y}_r + \tilde{R}\bar{Y}_{nr}) \\ &= \tilde{R}(\bar{Z}_{nr} - \bar{Y}_{nr}) \end{aligned} \tag{2.25}$$

This expression shows that imputation or substitution will result in a reduction in bias compared with Method 1 equal to the reduction (if any) of $|\bar{Y}_r - \bar{Y}_{nr}|$ to $|\bar{Z}_{nr} - \bar{Y}_{nr}|$ (Madow, Nisselson & Olkin 1983:26). The reduction in non-response bias, therefore, depends on the quality of the imputation or substitution procedures. Note that the non-response rate $\tilde{R}$ is not affected by imputation or substitution. Bearing in mind that ignoring missing data is equivalent to imputing the mean of respondents for non-respondents, it is clear that imputation is better than ignoring missing values provided that the imputed values provide a better estimate of $\bar{Y}_{nr}$ than $\bar{Y}_r$.

From the above can be seen that an effort to reduce non-response bias should be two-fold: (1) efforts should be made to reduce non-response to a minimum and (2) a statistical method should be identified to obtain an unbiased estimator $\hat{\bar{Y}}_{nr}$ of the mean among non-respondents (Lessler & Kalsbeek 1992:142). While efforts to reduce non-response (e.g., increased data collection efforts) will affect both $\tilde{R}$ and $|\bar{Z}_{nr} - \bar{Y}_{nr}|$, statistical methods of dealing with non-response, such as substitution and imputation, will affect only $|\bar{Y}_{nr} - \bar{Z}_{nr}|$ (Madow, Nisselson & Olkin 1983:26). In most cases, there will be a direct relationship between $\tilde{R}$ and $|\bar{Y}_{nr} - \bar{Z}_{nr}|$, i.e., increased data collection efforts will reduce $\tilde{R}$ and at the same time will result in improved estimates for $\bar{Y}_{nr}$ [8].

2.3.3. The Bias of Non-response: Stochastic View

Under the stochastic view of non-response, each element in the population has a random response probability $\varphi_j = E(R_j)$ where R_j is an indicator variable, defined as:

$R_j = 1$ if the j-th element, if selected, would respond

$R_j = 0$ if the j-th element, if selected, would not respond.

[8] Increased data collection efforts result in successive waves of data which can be analysed to obtain improved estimates of means for non-respondents (see Chapter 3).

2.3.3.1. Method 1: Do Nothing

Suppose a census is undertaken and the aim of the survey is to estimate the population mean $\bar{Y}$. Under the stochastic view, the observation actually used for the j-th population element may be expressed as:

$$R_j y_j \tag{2.26}$$

where the R_j are assumed to be mutually independent among population elements. If the missing data are ignored, an estimator of the population mean is:

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{j=1}^{N} R_j y_j \tag{2.27}$$

with bias:

$$\begin{aligned} Bias(\hat{\bar{Y}}) &= E\left[\frac{1}{N}\sum_{j=1}^{N}(R_j y_j - y_j)\right] \\ &= -\frac{1}{N}\sum_{j=1}^{N}(1-\varphi_j)y_j \end{aligned} \tag{2.28}$$

which shows that the size of non-response bias is related to the likelihood to respond.

If the deterministic view is assumed, (2.28) simplifies to:

$$\begin{aligned} Bias(\hat{\bar{Y}} \mid R_j) &= -\frac{1}{N}\left(\sum_{j=1}^{N_r}(1-\varphi_j)y_j \mid R_j\right) - \frac{1}{N}\left(\sum_{j=1}^{N_{nr}}(1-\varphi_j)y_j \mid R_j\right) \\ &= -\frac{1}{N}\sum_{j=1}^{N_r}(1-1)y_j - \frac{1}{N}\sum_{j=1}^{N_{nr}}(1-0)y_j \\ &= -\frac{1}{N}\sum_{j=1}^{N_{nr}} y_j \\ &= -\tilde{R}\bar{Y}_{nr} \end{aligned} \tag{2.29}$$

This expression shows that in a census, the absolute bias due to non-response is equal to the product of the non-response rate and the average "contribution" of the non-respondents.

2.3.3.2. Method 3: Imputation or Substitution

If the j-th population element fails to respond, an imputed value, $\hat{y}_{nr_j} = z_j$ may be used. If some imputation method is used under the stochastic view, the observation actually used for the j-th population element may be expressed as a random variable:

$$\hat{y}_j = R_j y_j + (1 - R_j) z_j \tag{2.30}$$

If $R_j = 1$, then $\hat{y}_j = y_j$. Note that z_j may be subject to error. The R_j are assumed to be mutually independent among population elements. The imputation error of using z_j to estimate y_j for the j-th population element is defined as:

$$E_2(z_j) - y_j \tag{2.31}$$

where E_2 denotes expectation over repeated applications of the imputation method under the same essential survey conditions. (Lessler & Kalsbeek 1992:140.)

Suppose the estimator of the population mean is:

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{j=1}^{N} \hat{y}_j \tag{2.32}$$

The bias of using $\hat{\bar{Y}}$ as an estimator of $\bar{Y}$ is:

$$\begin{aligned} E_1\left[E_2(\hat{\bar{Y}})\right] - \bar{Y} &= E_1 E_2\left[\frac{1}{N}\sum_{j=1}^{N}\{R_j y_j + (1 - R_j) z_j - y_j\}\right] \\ &= E_1 E_2\left[\frac{1}{N}\sum_{j=1}^{N}(1 - R_j)(z_j - y_j)\right] \\ &= \frac{1}{N}\sum_{j=1}^{N}(1 - \varphi_j) E_2(z_j - y_j) \end{aligned} \tag{2.33}$$

where E_1 denotes expectation over repeated simple random samples.

Expression (2.33) shows that the bias is jointly dependent on the likelihood of response and the quality of imputation (when necessary) associated with each population element. The absolute value of the bias will be a minimum (zero) when φ_j = 1 and a maximum when φ_j = 0 for any set of imputation errors that are either all positive or negative. The bias will be negligible when $E_2(z_j) \approx y_j$. (Lessler & Kalsbeek 1992:141.)

When the deterministic viewpoint is assumed (φ_j = 0 for $j = 1,2,\ldots,N_{nr}$ or φ_j = 1 for $j = 1,2,\ldots,N_r$), (2.33) becomes:

$$\begin{aligned} Bias(\hat{\bar{Y}} \mid R_j) &= \frac{1}{N}\left[\sum_{j=1}^{N_r}(1-\varphi_j)(z_j - y_j) \mid R_j\right] + \frac{1}{N}\left[\sum_{j=1}^{N_{nr}}(1-\varphi_j)(z_j - y_j) \mid R_j\right] \\ &= 0 + \frac{1}{N}N_{nr}(\bar{Z}_{nr} - \bar{Y}_{nr}) \\ &= \tilde{R}(\bar{Z}_{nr} - \bar{Y}_{nr}) \end{aligned} \tag{2.34}$$

which is the same as the result obtained in (2.25) where it was assumed that a census is conducted.

2.3.4. The Effect of Non-response Bias on Confidence Intervals

Suppose a 95% confidence interval is to be constructed for the population mean $\bar{Y}$ (see section 1.2.8). In the presence of non-response, the construction of the confidence interval will have to be based on the mean $\bar{y}_r$ of the respondents instead of on $\bar{y}$.

To examine the effect of non-response bias on this confidence interval, suppose the estimator $\bar{y}_r$ is normally distributed about its mean $E(\bar{y}_r)$ that is a distance $Bias(\bar{y}_r) = E(\bar{y}_r) - \bar{Y}$ from the population value $\bar{Y}$. The standard error of the estimate $\sqrt{V(\bar{y}_r)}$ is computed around the mean $E(\bar{y}_r)$ of the distribution instead of around the true mean $\bar{Y}$. (Cochran 1977:14.) The resulting interval I_r has the property that (Bethlehem & Kersten 1985:291):

$$P(\bar{Y} \in I_r) = \phi\left(1{,}96 - \frac{Bias(\bar{y}_r)}{\sqrt{V(\bar{y}_r)}}\right) - \phi\left(-1{,}96 - \frac{Bias(\bar{y}_r)}{\sqrt{V(\bar{y}_r)}}\right) \tag{2.35}$$

where $\phi(\cdot)$ is the standard normal distribution function. Table 2.2 gives values of $P(\bar{Y} \in I_r)$ for a range of values of $\frac{Bias(\bar{y}_r)}{\sqrt{V(\bar{y}_r)}}$.

Table 2.2 Effect of non-response bias on the confidence level (1 - α)

$\frac{Bias(\bar{y}_r)}{\sqrt{V(\bar{y}_r)}}$	$P(\bar{Y} \in I_r)$
0,02	0,9500
0,04	0,9498
0,06	0,9496
0,08	0,9492
0,10	0,9489
0,20	0,9454
0,40	0,9315
0,60	0,9079
0,80	0,8741
1,00	0,8300
1,500	0,6769

Adapted from Cochran, W.G. 1977. *Sampling Techniques*. New York: Wiley:14

From Table 2.2 can be seen that non-response bias has little effect on the confidence level provided that the bias is less than one-tenth of the standard error. At this point, the confidence level is 0,9489 instead of the 0,95 that it is presumed to be. As the bias increases further, the disturbance becomes more serious. At $Bias(\bar{y}_r) = \sqrt{V(\bar{y}_r)}$, the confidence level is 0,83 and not 0,95.

As a working rule, Cochran (1977:14) suggests that the effect of bias on the accuracy of an estimate is negligible if the bias is less than 10% of the standard error of the estimator. However, this rule is difficult to apply in the case of non-response bias, since it is usually impossible to find a guaranteed upper limit to $\frac{Bias(\bar{y}_r)}{\sqrt{V(\bar{y}_r)}}$.

2.4. CALCULATION OF RESPONSE RATES

2.4.1. Types of Response Rates

It has become standard practice in survey research in Europe and the USA - although not yet in the RSA - to report the achieved response (or non-response) rate. There is, however, a vast number of different ways in which this rate can be calculated. In general, it is calculated as a percentage rate which measures the number of respondents (or non-respondents) divided by some count of the number selected for participation in the survey (Lessler & Kalsbeek 1992:108). But variations occur in the way the numerator and denominator of the response rate are chosen. Lessler and Kalsbeek (1992:113) mention a study in which 40 survey research organisations were given the outcome data from a telephone survey and were asked to compute a response rate as they would report it. Twenty-nine (29) different rates ranging from 12% to 90% were produced, reflecting the lack of a standard definition of the response rate. On the other hand, different types of response rates may be used to measure the completeness of data in different ways and for different purposes (Groves 1989:140). The purposes of calculating response rates may generally be classified into two types:

1. *Statistical* - as a partial indicator of the extent of bias in the survey estimates
2. *Operational* - as an indicator of the success of field operations

In general, most rates calculated for statistical purposes are referred to as response rates while most rates calculated for operational or logistic purposes are called *completion rates*.

In this section, various response and completion rates are defined. The *sample response rate*, denoted as r, may be defined as the response rate that was actually achieved in the survey. The sample response rate is an estimator of the *population response rate*, denoted as R, i.e., $r = \hat{R}$. If the deterministic view is assumed, R can be interpreted to be the relative size of the respondent stratum in the population, i.e., $R = \frac{N_r}{N}$ (Nisselson 1983:112). If the stochastic view is assumed, R can be interpreted to be the proportion of elements *expected* to respond in samples of the given size or the proportion of the population that would have responded if the entire population were surveyed under the conditions of the sample survey (Madow, Nisselson & Olkin 1983:25; Platek & Gray 1986:25). Another interpretation is that R is the "estimated average response probability in the population" (Särndal *et al.* 1992:561).

In unequal probability samples, the population response rate is usually estimated by a weighted response rate which incorporates inclusion probabilities of sample elements, responding or not (Madow, Nisselson & Olkin 1983:26).

Some researchers calculate response rates based on some, but not all the elements selected for the sample: non-contacts or elements who were unable to be interviewed are excluded. Kviz (1977:265) warns against this practice:

> *Whatever the reason, if a nonresponse meets or is assumed to meet, all eligibility criteria for participation in a particular survey, that individual must be considered a potential respondent and counted in the calculation of the response rate. The fact that some individuals do not respond to the survey must be attributed to a failure on the part of the researcher in designing and executing the survey.*

Any sample response rate should therefore reflect the level of response among all *eligible* elements selected for the sample.

2.4.2. The Simple Response Rate, r

The most widely used response rate in epsem or self-weighting samples, called the *simple response rate*, is defined as the ratio:

$$r = \frac{n_r}{n} \tag{2.36}$$

where n_r is the number of *eligible* sample elements having responded to at least one item (Särndal *et al.* 1992:561) and n is the number of eligible elements selected in the sample. The simple *non*-response rate is defined as:

$$\tilde{r} = 1 - \frac{n_r}{n} \tag{2.37}$$

The simple response rate may be adjusted for use in all self-weighting sampling designs, for example, in stratified random sampling with proportional allocation (see section 1.3.2.1.2), it may be written as:

$$r = \frac{\sum_{l=1}^{L} n_{rl}}{\sum_{l=1}^{L} n_l} = \frac{n_r}{n} \tag{2.38}$$

where n_{rl} is the number of responding eligible sample elements in stratum l and n_l is the number of eligible sample elements in stratum l.

In multi-stage samples, the response rate should not be based on response in the ultimate stage only, but should take into account the level of response in all stages.

Suppose in stage I of two-stage element sampling the simple response rate among the PSU's is (see section 1.3.3.3):

$$r_I = \frac{n_{r_I}}{n_I} \tag{2.39}$$

where n_{r_I} is the number of eligible responding PSU's and n_I is the number of eligible PSU's in the sample. Suppose in stage II the response rate among elements is:

$$r_{II} = \frac{\sum_{k=1}^{n_I} n_{r_k}}{\sum_{k=1}^{n_I} n_k} \tag{2.40}$$

where n_{r_k} is the number of eligible responding elements in PSU k and n_k is the number of eligible elements in PSU k. The appropriate overall response rate for the survey is:

$$r = r_I \times r_{II} \tag{2.41}$$

The same principle holds in multi-phase samples, for example, in sub-sampling applied in follow-up for non-response (Madow, Nisselson & Olkin 1983:30) (see section 4.5). Suppose a sample of n elements is first selected and n_r respond. In a follow-up, a sample of m eligible elements is selected from the $n - n_r$ eligible non-respondents and m_r respond. The overall response rate should then be calculated as:

$$r = \frac{n_r}{n} \times \frac{m_r}{m} \tag{2.42}$$

Also consider the following example: In a survey of qualified remedial teachers who operate from schools in a certain province, the sample is selected in two phases. A simple random sample of 800 schools is first selected and asked to supply a list of remedial teachers associated with the school. In the second phase, a simple random sample of remedial teachers is selected from the lists obtained in the first stage. Suppose 640 schools respond so that $r_I = \frac{640}{800} = 80\%$. From information obtained from these responding schools, a second phase sampling frame is constructed listing 1 000 remedial teachers. A simple random sample of 200 remedial teachers is selected from the list and 190 respond, so that $r_{II} = \frac{190}{200} = 95\%$. These simple response rates provide an over-optimistic view of survey response, since the actual response rate is:

$$r = r_I \times r_{II} = 80\% \times 95\% = 76\%$$

2.4.3. Weighted Response Rates

When sample elements have unequal inclusion probabilities, it is preferable to calculate a *weighted* response rate, where the element counts are weighted inversely by the inclusion probabilities (Nisselson 1983:113). In a probability sample with known inclusion probabilities π_i for $i = 1, \ldots, n$, the weighted response rate is defined as:

$$r_w = \hat{R} = \frac{\sum_{i=1}^{n_r} \frac{1}{\pi_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i}} \tag{2.43}$$

Equation (2.43) simplifies to r in (2.36) if a self-weighting sample is used.

Groves (1989) and Nisselson (1983) do not seem to agree on the utility of the weighted response rate. According to Groves (1989:143), r_w estimates the likely response that would have been obtained if equal probability sampling had been used. Groves (1989:143) states (contradictory to Nisselson) that "only rarely ... would such an estimate be of use to the survey analyst" and instead of the single weighted response rate, he suggests calculating a separate response rate for each stratum or domain if the inclusion probabilities vary among the strata or domains.

If elements in separate strata have unequal inclusion probabilities (for example in disproportionate allocation in stratified random sampling - see section 1.3.2.1.2), the weighted response rate can be written as:

$$r_w = \frac{\sum_{l=1}^{L} \frac{n_{rl}}{\pi_l}}{\sum_{l=1}^{L} \frac{n_l}{\pi_l}} \tag{2.44}$$

Equation (2.44) simplifies to r in (2.38) if epsem sampling is used.

As an example, suppose 1000 sample elements are selected with unequal sampling fractions in two strata. Suppose the following values apply:

Stratum 1	$N_1 = 1000$	$n_1 = 500$	$n_{r_1} = 250$	$f_1 = \frac{500}{1000}$
Stratum 2	$N_2 = 9000$	$n_2 = 500$	$n_{r_2} = 450$	$f_2 = \frac{500}{9000}$

The simple response rate is:

$$r = \frac{250 + 450}{500 + 500} = 70\%$$

but if the unequal inclusion probabilities are incorporated, the following response rate is obtained:

$$r_w = \frac{\sum \frac{n_r}{\pi_i}}{\sum \frac{n_i}{\pi_i}} = \frac{\left(250 \times \frac{1000}{500}\right) + \left(450 \times \frac{9000}{500}\right)}{\left(500 \times \frac{1000}{500}\right) + \left(500 \times \frac{9000}{500}\right)} = 86\%$$

The relatively large difference between the simple and weighted response rates in this example is due to the large differences in size, sampling fractions and response rates between the two strata. (Särndal *et al.* 1992:561.)

2.4.4. Response Rates Based on Measures of Size

In some cases, weighted or unweighted unit *counts* are less informative than response rates based on known measures indicating the *size* of units (Nisselson 1983:113). For example, in a survey of business establishments, one may wish to express the extent of participation in the survey in terms of the *proportion of sales* represented by the respondents instead of the *number* of responses obtained (Lessler & Kalsbeek 1992:110). If M_i is the appropriate measure of size (for example annual sales, number of employees or crop yield) known for all units $i = 1, \ldots, n$, the weighted response rate (based on size measures) is:

$$r_M = \frac{\sum_{i=1}^{n_r} \frac{M_i}{\pi_i}}{\sum_{i=1}^{n} \frac{M_i}{\pi_i}} \tag{2.45}$$

This response rate should be used when the measure of size M is associated with the likelihood to respond: for example, larger establishments may be more likely to respond than smaller ones.

When the i-th unit is selected with *probability proportional to size,* that is $\pi_i = \frac{nM_i}{M}$ is the inclusion probability for the i-th unit, (2.45) reduces to the simple response rate r in (2.36):

$$r_M = \frac{\sum_{i=1}^{n_r}\left(\frac{M_i}{nM_i/M}\right)}{\sum_{i=1}^{n}\left(\frac{M_i}{nM_i/M}\right)} = \frac{\sum_{i=1}^{n_r}\frac{M}{n}}{\sum_{i=1}^{n}\frac{M}{n}} = \frac{n_r}{n} = r \tag{2.46}$$

In epsem samples, (2.45) is the proportion of total size measures in the sample represented by the respondents:

$$r_M = \frac{\sum_{i=1}^{n_r} M_i}{\sum_{i=1}^{n} M_i} \tag{2.47}$$

A response rate based on size measures is useful, since smaller establishments often have less potential for non-response bias in population estimates than larger establishments. When the main source of unit non-response is smaller establishments, r_M in (2.47) will be larger to reflect the more important contribution of large establishments to population estimates, even though r in (2.36) or (2.38) will be small.

Consider the following example of a simple random sample of business establishments. Average monthly sales for units in the sample are known approximately so that all business establishments can be classified into three size classes prior to sampling. Measures of size (average monthly sales) are calculated for each class by using respondent data and assuming that sales are approximately the same for respondents and non-respondents within each class.

Monthly sales (R-Million)	Size Class	Average Monthly Sales (R-Million)	Number of eligible units Responding	Sampled	Response rate (%)
< 1	Small	0,75	78	120	65
[1 ; 10)	Medium	6,50	60	80	75
≥ 10	Large	15,50	45	50	90
Total			183	250	73,2

A response rate based on the size of the units can be calculated as:

$$r_M = \frac{(78 \times 0{,}75) + (60 \times 6{,}50) + (45 \times 15{,}50)}{(120 \times 0{,}75) + (80 \times 6{,}50) + (50 \times 15{,}50)}$$
$$= \frac{1146}{1385} = 82{,}7\%$$

while a response rate based on unit counts gives:

$$r = \frac{183}{250} = 73{,}2\% .$$

Whereas r_M provides a larger response rate to reflect the more important contributions of larger establishments, r is smaller and does not discriminate between contributions by large or small establishments.

2.4.5. Completion Rates

In addition to, or sometimes instead of a response rate, various *completion rates* may be calculated with the main purpose to measure overall success in field operations. Whereas response rates are defined for eligible elements only, an advantage of completion rates is that they may be defined in a way that does not require a knowledge of the number of eligibles among the non-respondents (Madow, Nisselson & Olkin 1983:37). For example, a simple completion rate may be defined as:

$$c_1 = \frac{\text{Number of interviews with eligible households}}{\text{Number of households in sample}} \tag{2.48}$$

To illustrate the calculation of some completion rates, consider the categories in Figure 2.1, excluding for simplicity, the category of lost information. Assume that the remaining categories are collectively exhaustive and mutually exclusive and denote them (somewhat non-statistically) as follows:

NC = Number of *Non-Contacts*

I = Number of *Completed Interviews*

RF = Number of *Refusals*

U = Number *Unable*

NA = Number *Temporarily Unavailable*

Using these symbols, the basic completion rate in (2.48) may be written as:

$$c_1 = \frac{\mathrm{I}}{\mathrm{I + NC + RF + U + NA}} \tag{2.49}$$

Note that if there are no non-eligible cases among the non-contacts, (2.49) is equal to the *simple response rate* (2.36). However, determination of eligibility status is often a problem and NC will inevitably include some non-eligible elements. The only difference between the simple response rate in (2.36) and the completion rate in (2.49) is that the denominator of the latter is the sum of all eligible and non-eligible elements.

It is interesting to note that the "response rate" which is usually calculated in mail surveys, is actually a completion rate since the denominator inevitably includes some non-eligible sample elements.

Another type of completion rate, the *contact rate*, may be defined as the ratio of the number of contacted elements to the number of elements in the sample (Madow, Nisselson & Olkin 1983:30):

$$c_2 = \frac{\mathrm{I + RF + U + NA}}{\mathrm{I + RF + U + NA + NC}} \tag{2.50}$$

The contact rate may be used to assess how fully the sample elements were alerted to the survey. Madow, Nisselson and Olkin (1983:30) warn that the term "contacted" may have different meanings in different surveys and that it must therefore be defined very clearly.

The *co-operation rate* (the proportion of completed interviews obtained from those contacted) is defined as:

$$c_3 = \frac{\mathrm{I}}{\mathrm{I + RF + U + NA}} \tag{2.51}$$

It is a measure of interviewer ability to solicit participation once the sample element has been located (Groves 1989:141). Alternatively, the number of unable cases may be excluded from the denominator to obtain the rate:

$$c_3' = \frac{\mathrm{I}}{\mathrm{I} + \mathrm{RF} + \mathrm{NA}} \tag{2.52}$$

which measures interviewer success in obtaining responses from those *able* to do the interview. Alternatively, the denominator may also include all non-contact cases:

$$c_3'' = \frac{\mathrm{I}}{\mathrm{I} + \mathrm{RF} + \mathrm{NC} + \mathrm{NA}} \tag{2.53}$$

However, since NC may include some unable elements, c_3'' is an under-estimate of interviewer success to solicit participation from all able sample elements. (Groves 1989:141.)

Various other types of completion rates may be defined. These rates will differ according to the type of survey (census, telephone survey, face-to-face interview survey or mail survey) and which part of the fieldwork is being evaluated: location, solicitation, data collection or determining eligibility status (Lessler & Kalsbeek 1992:109). Whichever completion rate (or response rate) is used, it is important that a clear definition is provided when the survey results are published.

2.4.6. Rates Calculated by Reason for Non-Response

Non-response or non-completion rates calculated by reason for non-response are useful to measure bias due to various categories of non-response or to indicate success in data collection procedures. For example, the *refusal rate* may be defined as the proportion of (eligible) elements in the sample that refuse to respond. If eligibility status is uncertain, the refusal rate is actually a completion rate. Rates based on other reasons, such as sample element *unable* or *not found*, are similarly defined and may be used to detect problems in data collection. For example, a large proportion of non-response due to "element not found" may indicate that the frame is outdated or that more competent interviewers must be used. The *refusal conversion rate*, the proportion of sample elements who initially refused to provide an interview but later agreed to do so, may provide an indication of the interviewers' ability to persuade reluctant elements to respond.

2.4.7. Completion Rates for Household and Person Level

In household surveys, information is often desired for the household as well as for the individuals within the household. For example, one may be interested in the medical expenditure of the household as a unit but also in the medical history of individual members of the household. Non-response at the person level occurs when the eligible household member is absent and no other family member is able to respond on his or her behalf, i.e., a *proxy interview* with a knowledgeable family member cannot be obtained. Response rates (or completion rates) may therefore be desired for both the household level and the person level. (Madow, Nisselson & Olkin 1983:19.)

In order to define a completion rate for household and person level, the various categories in Figure 2.1 (excluding lost information) will be referred to. Assume that these categories are collectively exhaustive and mutually exclusive and distinguish between the categories at household and person level by the subscripts *h* and *p* respectively:

Household Level	Person Level (Known for responding households only)
NC_h = Number of Non-Contact Households	NC_p = Number of Non-Contact Persons
I_h = Number of Completed Household Interviews	I_p = Number of Completed Person interviews
RF_h = Number of Refusals for Full Household	RF_p = Number of Refusals for Persons
O_h = Other Non-Interviews for Households	O_p = Other Non-Interviews for Persons

Using these components, the household-level completion rate may be defined as (Groves 1989:144):

$$c_1(hh) = \frac{I_h}{I_h + NC_h + RF_h + O_h} \tag{2.54}$$

and the person-level completion rate as:

$$c_1(pp) = \frac{\mathrm{I_p}}{\mathrm{I_p} + \mathrm{NC_p} + \mathrm{RF_p} + \mathrm{O_p}} \tag{2.55}$$

The calculation of the person-level completion rate is complicated by the fact that the number of individuals in non-responding households may not be obtained, i.e., the number of individuals NC_p, RF_p and O_p are unknown for non-responding households. As a consequence, (2.55) is a completion rate that applies to persons in *responding* households only.

A solution to the problem is to estimate the average number of persons in the non-responding households, denoted as AVE (following the non-statistical notation adopted in this section!) and then multiply this estimate by the number of non-responding households. The person- level completion rate may then be calculated as:

$$\begin{aligned} c(pp) &= \frac{\text{Number of completed person interviews}}{\text{Number of persons in responding and non - responding households}} \\ &= \frac{\mathrm{I_p}}{\mathrm{I_p} + \mathrm{NC_p} + \mathrm{RF_p} + \mathrm{O_p} + AVE \times (\mathrm{NC_h} + \mathrm{RF_h} + \mathrm{O_h})} \end{aligned} \tag{2.56}$$

Groves (1988:144) suggests the average household size among responding households as a possible estimate of the average household size among non-respondents:

$$AVE = \frac{\mathrm{I_p} + \mathrm{NC_p} + \mathrm{RF_p} + \mathrm{O_p}}{\mathrm{I_h}} \tag{2.57}$$

The problem with this estimate is that responding household sizes often differ from non-responding household sizes (see Table 3.3).

2.4.8. Rates Calculated for Sub-classes

Response or completion rates are often calculated for sub-groups in the sample (for example, gender cross-classified with age) to determine whether the proportion of respondents in the sub-group is approximately equal to the proportion of the sub-group in the population (Madow, Nisselson & Olkin 1983:31). The population proportions within these sub-groups may be known from census figures or administrative sources. For example, suppose there is enough auxiliary information available to determine the gender and approximate age of non-respondents in the sample[9]. A low rate in the sub-group "male older than 40" with a high rate in the sub-group "female younger than 40" may indicate that working males are proportionally under-represented in the responding sample while housewives are proportionally over-represented. Additional data collection strategies, aimed specifically at working males, may then be devised.

Response or completion rates may also be calculated for *domains*. For example, if hostel-dwellers are a domain in a survey, "quality control" checks may be done by calculating completion rates at regular intervals during data collection to determine whether sufficient responses are obtained from hostel-dwellers to achieve the desired level of precision. Furthermore, separate completion rates may be calculated for interviewers, administrative sections or geographic areas. This will identify interviewers or areas that require additional data collection efforts. (Madow, Nisselson & Olkin 1983:31.)

2.4.9. Item Response Rates

In practice, *item* response rates can provide information on item response only for (unit) respondents. It is not always possible to know the number of eligible elements *in the sample* that should have responded to a particular *item* - but the number of (*unit*) *respondents* who should have responded to a particular item is generally known. Thus, it becomes practical to calculate item non-response rates for the entire eligible sample only if the number of non-responding elements that should respond to a particular item are known. Hence, an item response rate is usually calculated as the ratio of the number of eligible elements that do respond to an item to the number of *responding* eligible elements that should respond to that item. (Madow, Nisselson & Olkin 1983:32.)

In establishment or agricultural surveys, the *total* of an item or an auxiliary variable is often available for the population, for example, number of employees, sales or acreage. In such cases, item coverage rates (also called characteristic coverage rates) may be calculated for the item (Madow, Nisselson & Olkin 1983:32). The item or characteristic coverage rate is defined as:

[9] The gender of non-respondents may be obvious if the frame contains first names and the approximate age of non-respondents may be obtained from, e.g., neighbours.

$$C = \frac{\sum_{i=1}^{n_r} x_i}{\sum_{j=1}^{N} x_j} \tag{2.58}$$

namely the ratio of the sum of the (auxiliary) variable for all responding eligible units in the sample to the total of the variable for all eligible units in the population (or sometimes in the sample). The characteristic coverage rate expresses the extent to which the population (or the sample) is covered by the responses obtained (Madow, Nisselson & Olkin 1983:31). Note that if the auxiliary variable x is a measure of size of the unit, (2.58) (based on the sum of x for all eligible units in the *sample*) is equivalent to r_M in (2.47).

As with response rates based on size measures, the characteristic coverage rate is convenient when the main source of unit non-response is smaller establishments or farms. Because these smaller units *cumulatively* contribute little to the population total, they have less potential for bias. If the sample includes most of the large units, the item coverage rates will be large even with a relatively small unit response rate. (Madow, Nisselson & Olkin 1983:32.) For example: suppose in an agricultural survey the auxiliary variable x is the number of farm workers employed on farms in a certain district. Most of the large farms in the district are included in the sample and the non-response rate among large farms is 5% while among smaller farms it is, say, 60%. The total number of farm workers in the district is known from outside sources to be 10 000 while the total number of farm workers employed on responding farms is 9 000. The item coverage rate is:

$$C = \frac{9000}{10000} = 90\%$$

which provides a better indication than the simple response rate of the extent to which the responding portion of the sample covers the population.

2.5. INTERPRETATION OF RESPONSE RATES

2.5.1. Eligibility of Sample Elements

As was stated in section 2.4, only *eligible* elements should be included in response rate calculations. When eligibility status is a problem, Lessler and Kalsbeek (1992:112) suggest the calculation of two response rates: one assuming that all elements whose eligibility status is uncertain, are non-eligible (an upper bound) and the other assuming that these elements are eligible (a lower bound). There could be substantial variation between these bounds in surveys where eligibility status is unknown for a large proportion of the sample.

Using the notation from section 2.4.5 and assuming that eligibility status is known for all *contacted* elements in the sample (even the refusals), the upper bound for the response rate may be calculated as:

$$r_U = \frac{\text{I}}{\text{I} + \text{RF} + \text{U} + \text{NA}} \tag{2.59}$$

which is the same as the co-operation rate defined in (2.51). The lower bound may be calculated as:

$$r_L = \frac{\text{I}}{\text{I} + \text{NC} + \text{RF} + \text{U} + \text{NA}} \tag{2.60}$$

which is the same as the basic completion rate defined in (2.49).

Another possibility is to allocate elements of unknown status to eligible and non-eligible categories in the same proportion as among the elements of known status (White 1983:277). For example, suppose of all sample elements with known eligibility status, 10% are non-eligible and 90% are eligible. Then, for the purpose of response rate calculations, 10% of elements with unknown status are classified as non-eligible and the remainder are classified as eligible.

A comparison between response rates for mail and face-to-face interview surveys is complicated by the fact that in face-to-face interview surveys, non-eligible elements are easier to identify and exclude from response rate calculations than in mail surveys. Reported "response" (completion) rates in mail surveys are therefore often lower than what they would have been if all non-eligible elements were excluded.

Furthermore, all non-returns of questionnaires in mail surveys are often seen as refusals. According to Don A. Dillman, cited in Stopher and Sheskin (1981:252):

> *In face-to-face and telephone interviews a refusal is not considered as such until a contact is made. In mail studies, the opposite is assumed, that is, a nonresponse is a refusal until proven otherwise.*

2.5.2. Definition of the Target Population

Since eligibility is closely tied to membership of the target (or survey) population, it is important that clear definitions of the target population (and the survey population when there is a difference) are provided to show how eligibility is determined.

Consider the following example: in a survey to determine the opinion of consumers in a certain province about meat imported from Britain, the target population is defined as all non-vegetarian households in the province which are not self-sufficient for meat[10] (e.g., farmers who provide their own meat are excluded). According to this definition, all vegetarian households and some farmers in the province, will be non-eligible. However, if the target population is simply defined as all households in the province, such elements are eligible for the survey.

It is important also to consider the correspondence between the target population and the survey population when interpreting response rates. A survey with a high response rate but with large differences between the survey population and the target population, may be more biased than a survey with a low response rate but smaller differences between the target population and survey population.

Sometimes the survey population is especially defined in such a way that a higher "response rate" is obtained. Madow, Nisselson and Olkin (1983:34) caution against this practice:

> *To define a survey population by reducing or eliminating subsets having low response rates will lead to estimators having lower nonresponse rates in the survey population, but the estimators may have larger biases and possibly larger variances as estimators of characteristics of the target population. Inferences from the survey population to the target population may thus become very difficult.*

[10] This is an example of a survey where screening of the frame population is required to identify those elements who are eligible for the survey.

Continuing the example of consumer opinion above, suppose the province specified in the target population contains a large number of rural districts. If these districts are too costly to include in the survey or are expected to have high non-response rates, the *survey* population may then be defined as "all non-vegetarian households living in urban areas". If consumers living in rural towns differ in their opinions from those in urban areas, the effect of a high response rate in the survey population will be negated.

2.5.3. Imputation and Substitution Procedures

A frequent practice in dealing with non-response is to substitute for the non-respondents other available elements selected at random from the population or to impute information from other elements in the sample or population (see Chapters 4 and 6). It is important to note that substitution and imputation procedures do *not* affect response rates. (Madow, Nisselson & Olkin 1983:33.) As Deming (1953:34) states: "... there is no substitute for response"!

2.5.4. Accountability Tables

There are various reasons why response rates are published with survey results. One reason is to enable comparisons to be made between different surveys. Another reason is that when planning a new survey, achieved response rates of past surveys from a similar population may give an indication of the expected response rate in the new survey. It may also indicate an appropriate data collection technique to use and it may identify areas which need special attention when planning the survey. Madow, Nisselson and Olkin (1983:34) warn that to make meaningful comparisons between surveys, the definitions of the response rates should agree. In any survey, both the numerator and the denominator of the response rate should be *explicitly defined* and the population from which the sample is selected must be defined in sufficient detail.

In order to allow different users to calculate alternative rates in which they are interested, Madow, Nisselson and Olkin (1983:37) recommend that the information from which the rates may be calculated are given in the survey report in the form of "accountability tables" with the purpose "to account for the reduction from the number of elements in the sample to the number of elements for which responses are made".

The accountability table begins with the number of elements selected for the sample. All primary and secondary units in the sample are then described in detail in terms of eligibility and response status according to response outcome categories such as those in Figures 2.1, 2.2. or 2.3. Figure 2.4 is an example of an accountability table adapted from Madow, Nisselson and Olkin (1983:39).

Figure 2.4 Example of an accountability table

Occupied Household Survey: Face-to-face Interviews

Primary units:		
Total number of households in sample:		
	Occupied households Residential, vacant, non-seasonal Residential, vacant, seasonal Non-residential	
Interviews completed		
Non-interviews total		
	Refusal Not at home, repeated calls Illness Language problem	
Interview households not meeting requirements for secondary unit interviews		
Interview households meeting requirements for secondary unit interviews		
Secondary units:		
Total number of secondary units in sample		
Interviews completed		
Non-interviews		
	Refusal Not at home, repeated calls Illness Language problem	

2.5.5. Determining an Acceptable Response Rate

The question "what is an acceptable response rate?" is one that cannot easily be answered. Madow, Nisselson and Olkin (1983:5) admit:

> *Non-response rates should be "low", but whether 5%, 10%, 20% or some other percentage should be an upper bound for acceptable non-response rates depends on the survey objectives and is difficult to specify even for a particular survey.*

The following factors should be considered when deciding whether an achieved response rate is acceptable:

1. *The extent to which estimates for domains are affected by non-response.* In almost all surveys, estimates are provided for sub-groups (categories, domains or areas) of the population as well as for the entire population. Even if an overall non-response rate as low as 5% is obtained, estimates for some domains may still be seriously affected. It is therefore important that response rates are not only calculated overall but also for domains. (Madow, Nisselson & Olkin 1983:5.)

2. *The method of data collection used.* Since mail surveys normally have low response rates, 70% may be an acceptable level of response in mail surveys, but not in face-to-face interview surveys which are normally characterised by high response rates (Lessler & Kalsbeek 1992:116).

3. *Response rates from past surveys.* The achieved response rate in the current survey should be compared with past surveys from similar populations dealing with similar topics.

4. *Survey and operational characteristics.* Factors, such as the topic of the survey, respondent burden and the nature of elements in the target population should also be considered. For example, a response rate of 60% in a mail survey about deviant sexual behaviour may be considered "good" considering the sensitive nature of the topic. A lower response rate may have to be accepted in a survey requiring lengthy responses or record-keeping and a survey to determine entrepreneurial activities among illegal immigrants may just have to bear with some non-response!

5. *Respondent/non-respondent differences.* Although response rates are useful for describing participation in the survey, they say little about the *damage* caused by non-response (Lessler & Kalsbeek 1992:116). As was explained in section 2.3, the magnitude of non-response bias is determined by two factors: the non-response rate and the differences between respondent and non-respondent characteristics. The ultimate effect of non-response in a survey with a 90% response rate but large respondent/non-respondent differences may be more severe than in a survey with an 80% response rate but small respondent/non-respondent differences. In cases where non-response occurs approximately at random in the sample or at least in sub-groups of the sample (see Chapters 5 and 6), i.e., there are no differences between respondents and non-respondents, non-response (even a considerable amount) will not have a serious effect on survey results. The only effect of non-response in such a case is a reduction in sample size, since "the omission of a random part of a random sample leaves a smaller but still random sample" (Warwick & Lininger 1975:270).

6. *The aim of the research.* Some researchers, such as sociologists or psychologists, are primarily interested in the direction of relationships among variables and not in the estimation of population values (Kivlin 1965:325). Emphasis is on a certain phenomenon being studied and not on its distribution in the population. According to Goudy (1976:368), "the assumed impact of this [non-response] bias on variable relationships has been exaggerated". Other researchers have reported that non-response tended not to influence relationships between variables: Suchman (1962) has found that hypotheses that were tested, using biased and unbiased results, yielded the same conclusions. He concluded that where variables are independently related to the non-response bias, the use of biased or unbiased data will show the same inter-relationship between these variables. (See also Pavalko and Lutterman (1973:463), Mayer and Pratt (1966:637) and Hawkins (1975:462).)

2.5.6. Conclusion

According to Platek and Gray (1986:18) there is a need in survey research to monitor non-response rates and to ensure some degree of comparability between surveys, countries and survey organisations. This requires that surveyors (1) count the respondents and non-respondents according to type and reason and keep careful records of every sampled element as described in this chapter and (2) provide un-ambiguous definitions of the rates used. The lack of standardised definitions of response and completion rates in survey sampling literature is a problem, but according to Platek and Gray (1986:26):

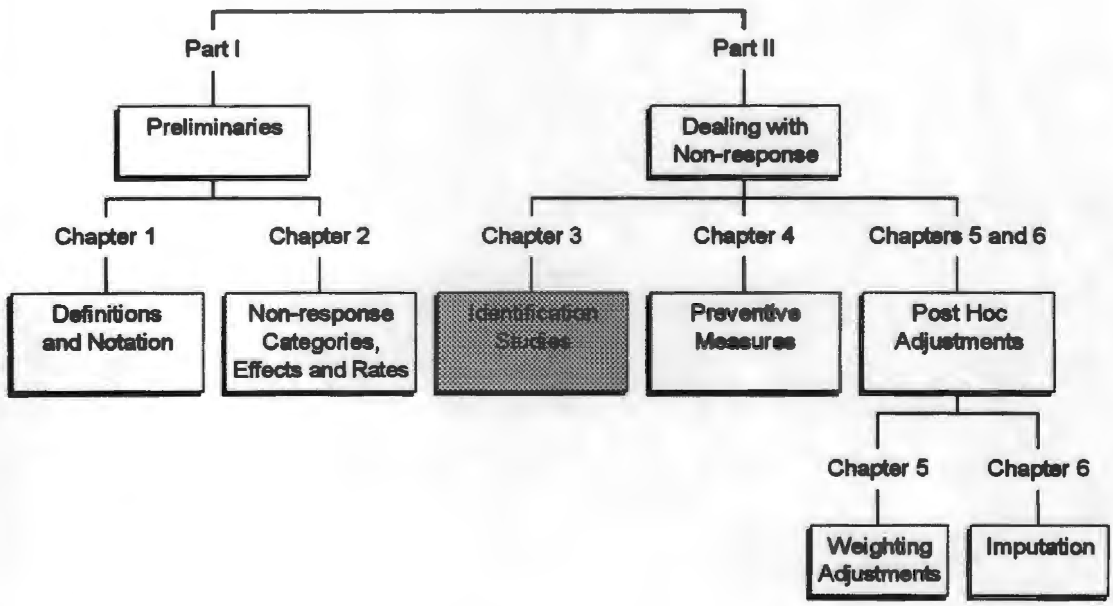
> *As long as the rates are unambiguously defined and appropriately applied in their analyses, standard definitions for all types of surveys and survey data gathering procedures may not be all that important. However, in each particular case, the rate should be carefully defined with clear demonstration of the purpose for which it is intended and the reason why it is adopted.*

PART II

DEALING WITH NON-RESPONSE

CHAPTER 3

IDENTIFICATION STUDIES



CHAPTER OUTLINE

CHAPTER 3

IDENTIFICATION STUDIES

If we want to discover what man amounts to, we can only find it in what men are; and what men are, above all other things, is various. It is in understanding that variousness - its range, its nature, its basis, and its implications - that we shall come to construct a concept of human nature, that, more than a statistical shadow and less than a primitive dream has both substance and truth.
Clifford Geertz, The Impact of the Concept of Culture on the Concept of Man

3.1. INTRODUCTION

In Chapter 2, the impact of non-response bias on survey estimates was shown to depend on two components: (1) the non-response rate and (2) the magnitude of the differences between respondents and non-respondents. A discussion of the various techniques that can be used to improve the response rate is deferred to Chapter 4. In the present chapter, various studies are considered which have aimed to qualitatively (or in some cases quantitatively) assess the impact of non-response on survey estimates by considering the differences between respondents and non-respondents in the survey. The title of the present chapter is the term given to such (mainly empirical) studies (Kalsbeek 1980:134).

After everything possible has been done to increase the response rate, the remaining non-response must be accepted as a *fait accompli*. Although efforts can be made to quantitatively estimate the residual biases due to non-response and adjust survey estimates accordingly, these efforts have not been very successful in the past (see section 3.5). The impact of the residual non-response on survey estimates can be assessed qualitatively by examining the nature of the differences that exist between respondents and non-respondents in the survey or, equivalently, by determining whether or not there is a relationship between response behaviour and the major survey variable(s). Specifically, if there are no differences between respondents and non-respondents in the population or if there is no relationship between response behaviour and the major survey variable(s), i.e., if the non-respondents constitute a *random* sample of the population, then non-response is considered to be "ignorable" irrespective of the response rate.

In order to compare respondent/non-respondent differences, the values of the survey variable(s) must be known for both respondents and non-respondents. However, the survey variable values are usually not obtainable for the non-respondents. In such cases, two possibilities exist:

1. Identify auxiliary variables which are highly correlated with the survey variables and with response behaviour and which are available for the entire sample. The existence of differences in auxiliary variable values between respondents and non-respondents is considered indicative of the existence of differences between respondents and non-respondents in the survey variables themselves.

2. Analyse characteristics of respondents at various points in the data collection process. The assumption behind this procedure is that sample elements who respond later in the survey indicate greater reluctance to participate and hence, stronger resemblance to non-respondents than those who respond early in the survey (Kalsbeek 1980:134).

Methods of identifying suitable auxiliary variables which can be used to study differences between respondents and non-respondents are discussed in section 3.2. Identification studies which analyse respondent values at various time periods in the survey are discussed in sections 3.3 and 3.4. The term "waves" refers to the sets of responses obtained after successive follow-up mailings, telephone reminders or call attempts. In section 3.5, some methods aimed at obtaining quantitative estimates of non-response bias are discussed. In the last section of this chapter, a selection of identification studies reported in literature are discussed briefly. The insinuation is not that their results apply rigidly to current or future surveys but that some value may be obtained from both the methodology used and the conclusions reached in these studies.

3.2. IDENTIFICATION OF AUXILIARY VARIABLES

An important aim of identification studies is to analyse the relationship (if one exists) between the survey variable(s) and response behaviour. This usually requires the identification of auxiliary variables that can be used in lieu of the survey variable(s) whose values are unobtainable for all non-respondents. Auxiliary variables commonly examined are socio-demographic characteristics such as age, ethnicity, gender and education. The auxiliary variables may, for example, be stratification variables but their values must be available for both respondents and non-respondents.

The identification of auxiliary variables which are related to response behaviour is not only important for studying the impact of non-response on survey estimates but is also necessary to define suitable weighting and imputation classes used to statistically adjust for non-response (see Chapters 5 and 6).

3.2.1. Graphical Exploratory Data Analysis

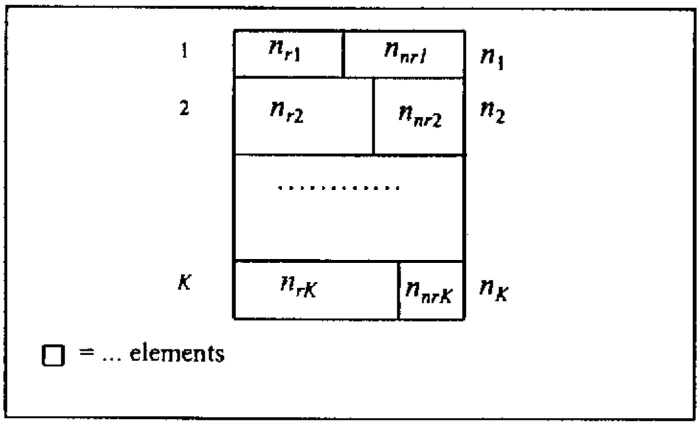
3.2.1.1. The Box Plot

According to Bethlehem and Kersten (1981a:6), it is useful to start the investigation of non-response with a graphical exploratory analysis since aspects of the data may emerge in this way which may otherwise be overlooked. One graphical exploratory technique which can be used is the *box plot*.

Suppose the auxiliary variable to be examined, x, consists of K categories: $1, 2, \ldots, K$. Although the box plot is usually applied in the case of a *single* (*categorical*) auxiliary variable, it is possible to incorporate more variables at the same time in the analysis (Bethlehem & Kersten 1981a:6). This is done by constructing a new variable for which the values are combinations of values of the original variables. For example, the variables *gender* (with values male and female) and *marital status* (with values married and unmarried) can be combined into a new variable with values male-married, male-unmarried, female-married and female-unmarried.

A box plot is constructed by drawing a rectangle with height proportional to the sample size and then dividing it into K layers to represent the categories of the auxiliary variable (see Figure 3.1). Suppose there are n_k elements in category k, of which n_{r_k} are respondents and n_{nr_k} are non-respondents ($k = 1, \ldots, K$). The height of each layer is drawn proportional to n_k, the number of sample elements in the corresponding category. Each layer is divided by a vertical line into a left-hand part, with area proportional to the number of respondents in the category (n_{r_k}) and a right-hand part, with area proportional to the number of non-respondents in the category (n_{nr_k}).

Figure 3.1 The Box Plot



A graphical exploration of the auxiliary variable and response behaviour in the survey is possible by examining certain aspects of the box plot:

1. The relative height of each layer which indicates the relative contribution of the various categories of the auxiliary variable in the sample.

2. The relative positions of the vertical lines with respect to the right-hand side of the box, indicating the proportions of non-response in each category.

3. The relative positions of the vertical lines with respect to each other, indicating the presence or absence of a relationship between response behaviour and the auxiliary variable. If all the vertical lines form approximately a straight line, there is no relationship or only a weak relationship between the auxiliary variable and response behaviour.

3.2.1.2. Example of the Use of a Box Plot

In order to illustrate the construction and interpretation of a box plot, suppose a mail survey is conducted to determine the level of job satisfaction among former students of a large university. The sampling frame is constructed from all final-year candidates registered with the university in 1990. Questionnaires are mailed to the permanent addresses supplied by these students while registered with the university. Suppose a final response rate of 70% is obtained after the initial mailing and two follow-up mailings. This is considered a satisfactory response rate, given the fact that the sampling frame is somewhat outdated. To determine whether there are potential differences between respondents and non-respondents in the survey, auxiliary information on the results of the final-year examinations is obtained from the university data files for the entire sample. The values of the auxiliary variable are classified into 5 categories; namely:

1 = results incomplete

2 = fail

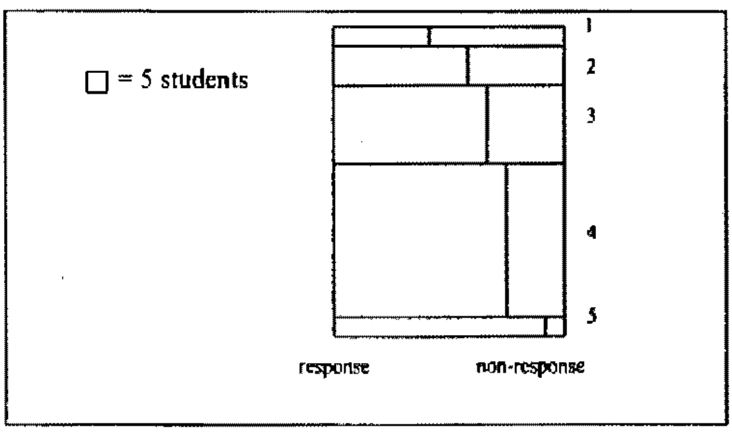
3 = pass - supplementary exam

4 = pass

5 = pass *cum laude*

The respondents and non-respondents in the survey are classified into the respective categories of the auxiliary variable and the following box plot is constructed:

Figure 3.2 ***Box plot showing relationship between final-year examination results and response behaviour***



The following conclusions can be drawn from the plot:

1. There are relatively fewer students who did not complete their degree, who failed or who passed *cum laude*. The majority of students passed either the major examinations or the supplementary examinations.

2. There is a relatively large proportion of non-response among students in the lower categories, i.e., students whose results were incomplete, who failed or who wrote supplementary examinations.

3. Since there seems to be a tendency of growing response as the results improve, the results of the final-year examinations seem to be correlated with response behaviour. To the extent that the results of the final-year examinations are

correlated with job satisfaction, the survey estimates based on respondent values will be biased.

It should be noted however, that a large proportion of non-response, especially in the "results incomplete" category, may be due to non-contacts instead of refusals: *firstly*, the sampling frame is not very current and *secondly*, some students in this category may have outstanding results because of reasons such as death or moving to another university or switching to a technikon. The interaction between the non-contact and refusal sources of non-response should be considered when applying corrective procedures. (See the study by Wilcox (1977) dealing with the interaction of various sources of non-response in section 3.6.9.)

3.2.2. CHAID Analysis and Logistic Regression

3.2.2.1. CHAID Analysis

Another method which can be used to examine possible correlations between auxiliary variables and response behaviour is CHAID, described by Kass (1980). CHAID is an off-shoot of AID (*Automatic Interaction Detection*) described by Morgan and Sonquist (1963). CHAID is a useful technique because it allows, in a more convenient way than the box plot, all specified auxiliary variables to be considered simultaneously. Furthermore, it can be used to identify the auxiliary variables and the cross-classifications of these variables, that have the strongest correlation with response behaviour. (Stoker 1986:154.)

A typical data set to be analysed by means of CHAID consists of a large number of vectors of auxiliary variable values (predictors) and a vector of values of the dependent variable. The dependent variable may be an indicator variable with the values 1 indicating response and 0 indicating non-response. CHAID partitions the data into mutually exclusive, exhaustive sub-groups that best describe the dependent variable, (e.g., response behaviour). At each partition, the sub-groups are based on the categories of the auxiliary variable with the most significant Chi-square value selected from the pool of possible predictors. The selected predictors may then be used in further analyses (e.g., for the definition of appropriate *weighting* and/or *imputation classes* for non-response adjustment (see Chapters 5 and 6)), prediction of the dependent variable or, in the place of the total set, in subsequent data collection. (Kass 1980:121.)

CHAID has a number of disadvantages:

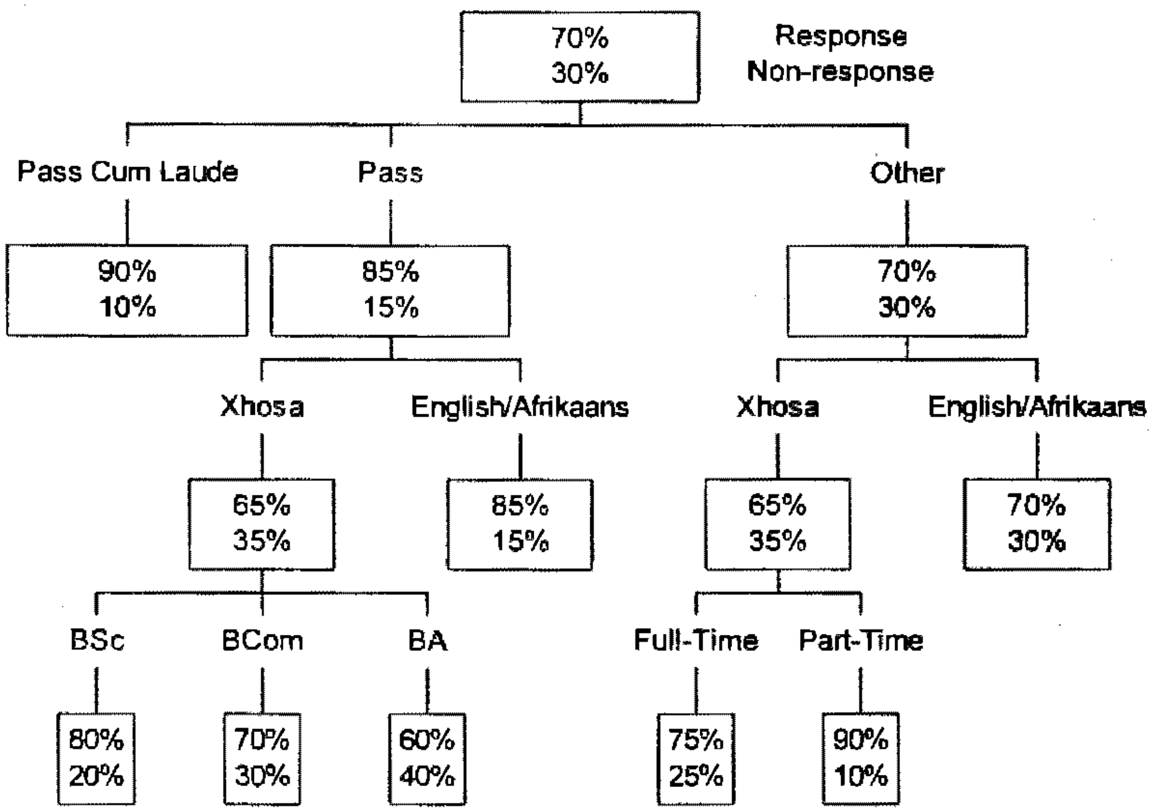
1. The specific sub-groups identified by CHAID may be complex and rather awkward to work with in practice (Chapman 1976:248).

2. If the sub-groups are used to define weighting or imputation classes, some classes may contain a very small number of respondents which may lead to an increase in the variance of weighted estimators. Kalton (1983a:64) allows the minimum number of respondents per sub-group to be 20 to 25. If this minimum number cannot be attained, categories of one or more auxiliary variables must be pooled until this requirement is met. The categories of those auxiliary variables showing the weakest association with response behaviour should be joined. (Stoker 1988:45.)

3. CHAID actually requires the use of a self-weighting design. The use of weighted data in CHAID leads to the formation of more "significant" splits in the data than is actually the case. (Stoker 1988:44.)

Consider a hypothetical application of CHAID to the example described in section 3.2.1.2). Suppose the data set consists of the response indicator variable (response/non-response) and amongst others, the following auxiliary variables as possible explanatory variables: final-year examination results (described in section 3.2.1.2), home language (Afrikaans, English, Xhosa), degree (BCom, BA, BSc) and whether the candidate was a full-time (FT) or part-time (PT) student. The outcome of CHAID may be as in Figure 3.3:

Figure 3.3 Hypothetical CHAID-analysis of response behaviour



Suppose the number of elements in the categories "Pass Cum Laude", "Pass English/Afrikaans" and "Other English/Afrikaans" were less than the minimum number specified in the analysis so that no further partitionings of these sub-groups were made. These categories are referred to as terminal nodes. From the analysis can be seen that the variable "exam results" has the strongest correlation with response behaviour, followed by the variables "language" and "degree" and "full-time" or "part-time". Weighting and imputation adjustments for non-response can now be based on the sub-groups identified in Figure 3.3.

3.2.2.2. Logistic Regression

Another technique which can be used to identify auxiliary variables that are correlated with response behaviour is logistic regression. Logistic regression is suitable for analysing a data set with an indicator dependent variable (e.g., $Y = 1$ if "response" and $Y = 0$ if "non-response") and some explanatory variables (i.e., auxiliary variables thought to be correlated with response behaviour). By fitting a logistic regression model to the data, auxiliary variables with coefficients significantly different from zero, can be identified as being correlates of response behaviour.

3.3. WAVE ANALYSIS

Probably the most effective method of increasing the response rate is to call-back repeatedly on non-respondents or to send reminder letters by mail (see Chapter 4). The result is "waves" of responses that can be utilised to analyse differences between respondents and non-respondents. The rationale for using waves in identification studies is that sample elements who respond on the t-th wave indicate greater reluctance to participate (and thus stronger resemblance to non-respondents) as t increases (Kalsbeek 1980:134).

In the place of data on hard-core non-respondents, survey analysts performing wave analyses use data obtained from *respondents* who refused to participate on the initial contact but were later persuaded to participate. The "non-response" bias at various stages of the survey is estimated as the differences between the estimates obtained up to a particular wave and the final estimates obtained after all waves. For example, analysts may estimate what the non-response bias *would have been* if only two call-backs had been made. The standard of comparison of estimates at various stages is thus the final estimates obtained from the survey which are themselves subject to non-response bias due to hard-core non-response. Except where the follow-up virtually eliminates all hard-core non-responses, these studies fail to estimate the true bias due to non-response. These studies are aptly referred to by Scott (1961:157) as "studies of *early* versus *late* response bias" and by Hawkins (1975:464) as "studies of reluctant respondents".

Another short-coming of many identification studies using the wave analysis technique is their emphasis on distributional bias of the auxiliary variables - usually demographic characteristics of respondents and non-respondents - rather than on the impact of non-response on the actual survey estimates. An exception is wave analyses which attempt to quantitatively assess the non-response bias of survey estimates by extrapolating the values of the survey variable to hard-core non-respondents.

3.4. EXTRAPOLATION METHODS

In this type of wave analysis, data from successive response waves are analysed, usually on a cumulative basis. The respondents to successive follow-ups are viewed as distinct categories forming a continuum which range from willing to reluctant respondents. Each wave is seen to probe deeper into the core of non-respondents so that by extrapolating over successive waves, estimates of the characteristics of hard core non-respondents may be obtained. (Filion 1975:484.)

An estimate of the population value is obtained at each wave, usually based on the cumulative responses up to date. This sequence of estimates is then fitted to a regression model and the fitted model is used to estimate what the population value would be for a 100% response rate. The independent variable (x) used in these models may be a response variable indicating the degree to which the respondent is assumed to resemble a non-respondent (e.g., 1 = willing respondent, 2 = reluctant respondent, 3 = unwilling respondent) or it may be a measure of the resistance of the sample elements to respond (e.g., the time to completion of the interviews: 1 call, 2 calls, 3 calls). Another example of an independent variable that may be used is the cumulative response rate at each call. The dependent variable (y) is usually the estimate of the population value obtained by pooling data from the start of data collection through the point determined by x. (Lessler & Kalsbeek 1992:173.) Of course, these regression models will be less successful to the degree that "hard-core" non-respondents differ from "late" responders. Furthermore, the final response rate determines to what extent the model can be used for extrapolation. For example, if the final response rate is 30%, serious doubts can be cast on the ability of the model to extrapolate to a 100% response rate.

The definition of x determines the manner in which the fitted model is used for extrapolation. When x is the cumulative response rate, the extrapolated estimate is for the y variable in the case of complete response. However, when x is the wave number, number of the call attempt or length of time to completion of the interview, the object of extrapolation is less clear. In such a case, one must make some arbitrary choice of the value of x (e.g., 10-th wave, 2 months) that will adequately predict the value of the y variable for hard-core non-respondents. (Lessler & Kalsbeek 1992:174.)

An early example of an extrapolation method is the log-linear model used by Hendricks (1949) in a mail survey of 3241 fruit growers in North Carolina:

$$y = \alpha x^{\beta} \tag{3.1}$$

This model was fitted to the data given in Table 3.1. In this example, y represents the average number of trees per farm and x represents the number of the wave on which the response was obtained.

Table 3.1 Results from repeated mailings to fruit growers

Mailing	Number of questionnaires returned	Average number of trees per farm
1	300	456
2	543	382
3	434	340
Known Population Value	3241	329

Source: Hendricks, W.A. 1949. Adjustment for bias caused by non-response in mailed surveys. *Agricultural Economics Research*. 1949:52.

The predicted value obtained from the model (3.1) was 329,9 which is approximately equal to the known population value of 329. Hendricks (1949:55) warns that data from at least three mailings are necessary before the method can be applied.

Filion (1975, 1976) used a *linear* regression model in a mail survey of Game Bird Hunting Permit purchasers in Ontario to express various survey variables as a function of the cumulative response rate after each wave of replies. The model used by Filion (1975, 1976) is:

$$y = \alpha + \beta x \tag{3.2}$$

where y is the observed value of the survey variable(s) based on the responses up to a given wave and x is the cumulative response rate up to a given wave.

In the Game Bird Hunter survey, three waves of responses yielded a total response rate of 79%. The estimate of the proportion of potential hunters below age 40 obtained from the linear regression model was 45%. This estimate is equal to the known population value of 45%. Furthermore, the model estimated the proportion of hunters residing in rural areas as 87% while the true population value was 85%. (Filion

1975:490.) Despite the degree of success which has been obtained by using extrapolation methods, fitting a regression model to 3 data points only may result in very unstable estimates.

Ognibene (1971) conducted a telephone (directory) survey of men in the New York metropolitan area, followed three months later by a mailed questionnaire which repeated items asked in the telephone survey. The responses obtained to the telephone survey were used as the "population values" to compare characteristics of the respondents and the non-respondents to the mail survey. A 34% response rate was obtained in the mail survey. Ognibene (1971) compared the feasibility of various models relating the cumulative response rate to the survey variables in this survey and concluded that the most useful results were obtained from the hyperbolic model:

$$y = \alpha + \frac{\beta}{x} \tag{3.3}$$

In the Health and Nutrition Examination survey, Chapman (1976) computed the mean of a survey variable among respondents requiring only one call, among those requiring two calls, etc., and attempted to fit a regression model to the data. Chapman (1976:249) concluded that it was not possible to determine a general trend between the mean responses at each call. However, according to Scott (1961:162), it is more appropriate to treat the waves cumulatively rather than individually as was the case in the study by Chapman. *One* reason is that the choice of independent variable to represent the non-respondents when treating the waves individually is somewhat problematic (5 calls? 10 calls?); *another* is that by treating the responses cumulatively, more stable estimates of the survey variable in the case of complete response may be obtained.

Hochstim and Athanasopoulos (170:76) concluded in the same manner that there was insufficient evidence of a trend in responses over the successive waves in their study. Scott (1961) found little evidence of a strong relationship when the y variable is any one of several demographic characteristics, although some empirical findings are mentioned which suggest that estimates obtained by extrapolation may sometimes be useful.

Armstrong and Overton (1977) used extrapolation methods on data from 11 studies (with a total of 112 items) that each had three response waves. Extrapolation was based on *two* waves only while the cumulative responses up to the third wave were used as "population values". (The use of only two data points for extrapolation may lead to very unstable estimates!) The results showed that the "error" from extrapolation (the difference between the extrapolated value and the cumulative responses from three waves) was substantially less for these items than the "error" using no extrapolation (the difference between the results from two waves and the cumulative responses from three waves). A serious limitation of this study is that predictions were made for third wave responses while, in practice, one would be extrapolating to a 100% response rate.

Clearly, the utility of extrapolation in dealing with survey non-response is not firmly established (Lessler & Kalsbeek 1992:174). In most studies, the strength of the relationship between measures of the times of completion and the survey variable is questionable. On the other hand, the extrapolation approach has some advantages (Lessler & Kalsbeek 1992:174; Filion 1975:492):

1. It is relatively simple and inexpensive
2. It requires no auxiliary information from external sources which may themselves be subject to non-response bias and/or measurement error
3. It utilises the *survey* variables and not *auxiliary* variables whose estimation is not the object of the survey
4. The results of extrapolation can be obtained without the expense of non-respondent sub-sampling or complicated weighting schemes

Lessler and Kalsbeek (1992:175) conclude that extrapolation methods cannot be used widely until measures of x which better predict values of the y variable can be found. Many extrapolation methods fail because explanatory variables that represent the respondents' likelihood of participation are not available. For example, the number of calls required to obtain a response in an interview survey may reflect only the respondents' availability, not willingness or ability to participate. In mail surveys, the wave in which response occurs and the time to completion may only indicate the respondents' tendency to procrastinate or forget. Presumably, an ideal response measure would reflect all the possible reasons for non-response. (Lessler & Kalsbeek 1992:175.)

3.5. ESTIMATING NON-RESPONSE BIAS

In Chapter 2, an expression for the non-response bias of $\bar{y}_r$ as an estimator of $\bar{Y}$, was derived as:

$$\begin{aligned} Bias(\bar{y}_r \mid n_r) &= E(\bar{y}_r \mid n_r) - \bar{Y} \\ &= \tilde{R}(\bar{Y}_r - \bar{Y}_{nr}) \end{aligned} \tag{3.4}$$

Of course, non-response bias would not be a problem if $Bias(\bar{y}_r \mid n_r)$ were known. In such cases, an *unbiased* estimator for $\bar{Y}$ could still be constructed as:

$$\hat{\bar{Y}} = \bar{y}_r - Bias(\bar{y}_r \mid n_r) \tag{3.5}$$

Unfortunately, the size of the non-response bias is usually unknown. An estimator of the non-response bias would require:

1. An unbiased estimator of $\tilde{R}$, the population non-response rate. $\tilde{R}$ may usually be estimated by the sample non-response rate $\tilde{r}$.

2. An unbiased estimator of $\bar{Y}_{nr}$. $\bar{Y}_{nr}$ may usually be estimated by $\bar{y}_{nr}$ which, in turn, needs to be estimated in some way from the *respondent* values in the sample.

Possible methods to obtain quantitative estimates of the non-response bias (3.4) include:

1. Consider sample elements who respond "late" in the survey as being a random sample of elements who do not respond at all. The mean of the late responders is used as an unbiased estimate of $\bar{y}_{nr}$.

2. If, through some extraordinary means, a *random* sub-sample of non-respondents can be convinced to participate, estimates of the non-response bias may be obtained by using the mean of the sub-sample as an unbiased estimator of $\bar{Y}_{nr}$.

Sub-sampling as a method of dealing with non-response is discussed in Chapter 4.

Methods of obtaining *quantitative* estimates of non-response bias have not been very successful so that, in general, one is left in the position of "relying on some guess about the size of the bias, without data to substantiate the guess" (Cochran 1977:361). On the other hand, it is better to rely on such tentative estimates than to ignore a potentially troublesome problem in the hope that inattention will cause it to go away.

Cochran (1977:361) suggests that information may be obtained from an external source so that *bounds* can, at least, be placed on $\bar{Y}_{nr}$. This technique may be useful in estimates of *proportions* but it is less useful with continuous variables since the only bounds that can be assigned with certainty to an unknown continuous variable are often so wide as to be useless (Cochran 1977:361). Cochran (1977:361) gives a method to construct confidence limits for the population proportion by assuming, for the lower limit, that the population proportion for non-respondents is 0 and, for the upper limit, that the population proportion for non-respondents is 1. However, even these limits are "distressingly wide" unless the non-response rate is very small.

In section 3.6, the results of a few empirical identification studies which have aimed to *qualitatively* assess the differences between respondents and non-respondents in specific surveys, are discussed. It is important to note that the few demographic or other correlates of response behaviour that have been identified may be significant for explaining non-response in some surveys but not in others.

3.6. SOME EMPIRICAL RESULTS

3.6.1. Study by Donald (1960)

In a mail survey of the League of Women Voters in the USA, Donald (1960) found that response rates tended to be high for members who were actively involved in the activities of the League.

Three follow-ups were used in the survey: two reminder letters sent by mail followed finally by a telephone call to the remaining non-respondents, yielding a final response rate of 77,3%. Provision was made in the questionnaire for the planned analyses of response behaviour by including questions on involvement in the activities of the League. Furthermore, the telephone follow-up was designed not only to increase the response rate, but also to ask refusers brief questions concerning the extent of their participation in the activities of the organisation. Demographic characteristics of respondents to the first and second wave were compared with those of respondents to the third and fourth waves. Information on the level of involvement that was obtained for some of the hard-core refusers to the main survey was also compared with that of the respondents. At least half of the non-respondents did not actively support the League.

A significant relationship was obtained between response elicitation and member involvement in the organisation. The higher the involvement, in terms of active participation, knowledge and understanding of the organisation and loyalty to it, the fewer the stimuli required to induce a response.

3.6.2. Study by Hilgard and Payne (1944)

Hilgard and Payne (1944) studied a household survey of consumer requirements in the USA in which records were kept of the number of calls required for each household before an interview was obtained[1]. Interviews obtained on later calls were analysed to give a picture of the kind of people less often home and to provide a basis for estimating the bias which would result if they were not reached in the survey. Some of the more significant results of the study are summarised in Table 3.2.

Table 3.2 Characteristics of Urban Households Interviewed by Call Number

Characteristics of households		Percentage of households interviewed on: 1st call	2nd call	3rd or later call	All households interviewed
EMPLOYMENT:	Not employed outside home	78,2	57,8	46,4	69,1
	Employed outside home	21,8	42,2	53,6	30,9
CHILDREN:	Has children under 2 years	17,2	9,5	6,2	13,9
	Has no children	45,2	55,7	61,5	49,8
HOUSEHOLD SIZE:	1 person	6,3	13,1	15,2	9,1
	5 or more persons	23,6	17,7	12,7	20,7
	Average size	3,56	3,11	2,84	3,35

[1] The archaity of this study is not believed to distract from the relevance of the results to the study of differences between respondents and non-respondents in current surveys. This is further evidenced by a number of references to this study in (more current) statistical literature, e.g., Särndal *et al.* (1992:565).

A number of conclusions can be made from Table 3.2. *Firstly*, it can be seen that respondents employed outside the home were much harder to reach than those not employed outside the home. Although only 21,8% of respondents reached on the first call were employed outside the home, 53,6% of respondents reached on the third and later calls were employed outside the home.

Secondly, households with young children were easier to reach than households with no children at all. If no follow-ups were made and only first call results were used, the results would consist of too many respondents with young children and too few with no children.

Thirdly, single-person households seemed to be the most difficult to reach while members of large families were easier to contact. The average household size of first-call respondents was 3,56, while the average household size of respondents on third or later calls was 2,84. The researchers gave no results on the number of refusers or on the extent of possible bias caused by refusals.

3.6.3. Study by Mayer and Pratt (1966-67)

In a mail survey to explore the economic and psychological consequences of personal-injury motorcar accidents in Michigan, Mayer and Pratt (1966-67) were able to compare demographic characteristics of respondents and non-respondents to the survey. A comparison could also be made between non-respondents who refused to respond and those who could not be contacted. Such comparisons were possible because auxiliary information on characteristics such as age, sex, race and occupation was available in police reports for all sample elements. The survey involved one mail follow-up letter and a final telephone call to the remaining non-respondents. The final response rate was 74,4%.

Significant differences were obtained for race and occupation categories over the three response waves. Results taken from the first wave alone would have been seriously biased towards whites and similarly towards professionals, self-employed businessmen, managers and clerical and sales personnel, whilst craftsmen and foremen were under-represented in the first wave. Mayer and Pratt found, however, that the next two waves tended to eliminate the first-call bias. Furthermore, they found that estimates based solely on first-wave results were very similar to the cumulative three-wave estimates and that the cumulative three-wave distribution was "very similar" to the known distribution of characteristics in the entire sample. The researchers concluded that, although there was a 25,6% non-response rate, the non-response bias was negligible for the characteristics considered in this particular survey.

When the researchers compared the characteristics of *refusers* and *non-contacts*, statistically significant differences were obtained for sex, age, race and occupation. Those who could not be contacted tended to be male, younger than 24, non-white and tended to have lower occupational skills.

In this study, results for a given survey are compared with known values for the population in order to obtain an estimate of the differences between respondents and non-respondents in the survey. However, because the "known" population values come from a source external to the survey, differences may occur between the population and survey values as a result of measurement errors. It is also not always possible to determine whether the data from the external source are free from non-response bias (Armstrong and Overton:1977:397).

3.6.4. Study by O'Neil (1979)

In a telephone survey of households in Chicago, discussed by O'Neil (1979), up to 20 calls were made at staggered times to some households. A proportion of 86,8% of the sample elements who were contacted eventually completed the interview. Sample elements who refused to be interviewed on the first contact, were mailed a persuasion letter and re-contacted by telephone. The refusal conversion rate was 44,4%. In order to compare the characteristics of respondents and non-respondents, those who responded to the first call were compared with those who responded after having refused on the first call.

Highly significant occupational differences were found between the two groups in this survey: white collar workers were less likely than blue collar workers to refuse an initial interview. Patterns for college and graduate students paralleled those for the white-collar workers, while trends for high-school students resembled those for blue-collar workers. Furthermore, those in the lower income categories, the less educated, whites and persons over 65 and under 19 years were more likely to refuse an interview.

3.6.5. Study by Gannon, Nothern and Carroll (1971)

A survey was conducted in Washington DC to determine the attitudes of employees of a supermarket chain about various job conditions (Gannon, Nothern & Carroll 1971). The final response rate to the survey was 63%. (No information is given by the researchers about the use of follow-up procedures.) The researchers compared characteristics of respondents and non-respondents to the survey by consulting the administrative records of the company. They found that a low response rate was associated with lower levels of education, single status, male status, being younger or older than the middle-aged group (between 30 and 49 years of age) and ranking in the bottom fifth of all employees on supervisory ratings. Based on the information in the administrative records of the company, the researchers concluded that respondents in this survey seemed to be more stable, older and more effective employees than non-respondents.

3.6.6. Study by Dunkelberg and Day (1973)

In a study of the 1967 Survey of Consumer Finances in the USA, Dunkelberg and Day (1973) attempted to provide empirical evidence of the relationship between the number of call-backs and bias in the distributions of population characteristics. The rate at which the distributions of selected characteristics converged on their *estimated* population values, based on the results after 7 or more calls, was empirically estimated and compared across various sub-groups of the sample. The final response rate in the survey was 82%.

The results of the study showed that nearly 80% of the final number of respondents over 64 years of age were contacted in the first and second calls, while the corresponding percentages for the other age groups were considerably lower. A disproportionate fraction of the less-educated and low-income families were reached early in the interviewing process but, on the other hand, less than half of all respondents living in the twelve largest central cities in the USA had been reached by 2 calls. Respondents living in the twelve largest central cities were therefore the most difficult to reach, while retired respondents or those employed in the home were the easiest to reach.

The researchers performed AID analysis using 16 possible predictors, some of which were city size, sex, marital status, education, race, type of dwelling unit, age, family income, employment status, family size, house value and occupation. The AID analysis suggested that the most important characteristic explaining the variation in the number of calls required was city size of residence. The joint effect of the 16 characteristics selected for the AID analysis explained only 13% of the variance in the number of calls required. The researchers believed that variables, such as the day of the week and the time of the call, should also have been added to the pool of possible predictors since these variables may have explained a significant amount of variation in the number of calls required.

When estimating the rate of convergence, the researchers found that overall, the distributions of the variables studied converged rather quickly to their "true" sample values (the values obtained after 7 or more calls): most were found to converge after two or three call-backs. The researchers admit:

> *The final distribution in the survey is used as the final or proper distribution to be achieved. A more satisfactory procedure might be to use exogenous estimates of the distributions (from census data for example). This requires careful matching of sampling procedures and frames as well as the definitions of the interview unit and procedures used to reach the unit.*

3.6.7. Study by Hawkins (1975)

Hawkins (1975) compared the characteristics among respondents in a household survey of adults in Detroit according to the number of calls required before an interview was obtained. An average of 3 calls per household were required but a total of 17 calls were required to some households to yield the final response rate of 72%.

The (cumulative) distributions of various demographic variables were compared (1) between the first and last calls and (2) between the sixth and last calls (see Table 3.3). The analyses showed that some group of respondents were initially greatly under-represented, whilst others were over-represented. The differences in the distributions between the first and the final calls were most severe for the white-collar group, the college-educated group and Jewish respondents who were all under-represented in the initial calls and for the lowest income group, the aged and housewives and students who were all over-represented in initial calls. By the sixth call, many of these differences had diminished greatly and an almost complete convergence to the final sample values could be observed. However, even between the sixth and final call, the unemployed and retired and the over-65 age group were appreciably over-represented whilst blue collar workers, Jews and blacks were appreciably under-represented. For most variables, at least three to six calls were required before a distribution approximating the *final* sample distribution was achieved.

Table 3.3 *Cumulative distributions (%) of respondent characteristics*

Respondent Characteristics		Number of Calls 1	6	10+
SEX	Male	33,6	40,9	41,1
	Female	66,4	59,1	58,8
RACE	Black	17,7	19,5	21,4
RELIGION	Jewish	3,8	6,2	6,5
EDUCATION	College graduate or more	8,0	13,6	13,5
OCCUPATION	White Collar	15,9	31,2	31,7
	Housewife and student	44,3	33,1	31,9
	Unemployed and retired	10,6	9,9	9,3
AGE	Over 65 years	12,4	9,7	9,0
FAMILY INCOME	$0-$5 999	23,7	18,9	18,6

3.6.8. Other Studies

Goudy (1976) found that those in the upper income and education levels tended to respond earlier than other groups. Hochstim and Athanasopoulos (1970) found that non-respondents were older and more likely to be white, male and skilled workers. In the latter study, information about the non-respondents was obtained from a previous survey.

3.6.9. Study by Wilcox (1977)

There are many conflicting results in literature about the characteristics of respondents and non-respondents of which ample proof can be found in the few studies discussed in this section. For example, while some researchers have found refusers to be of lower economic status and less well-educated than respondents (O'Neil 1979) (see section 3.6.4), other researchers (Dunkelberg and Day 1973; Hawkins 1975) (see sections 3.6.6 and 3.6.7) suggest the opposite profile for hard-to-reach sample elements: those with college degrees and higher incomes require more call-backs to be found whereas less well-off respondents are easier to find.

Wilcox (1977) suggests that a possible reason for these conflicting results is the interaction between two sources of non-response; namely, refusal and not-at-home if they are related to the same demographic items. Wilcox (1977:595) identifies several potential types of interaction. For example, suppose a specific group of respondents in the sample is both difficult to find at home and likely to refuse once they are found. In this case, the biases of the not-at-home and refusal sources of non-response would tend to reinforce each other and this group would be under-represented in the sample. On the other hand, if a particular group is both easy to find and co-operative, it would be over-represented in the results.

The two sources of bias could also work in opposite directions. For example, if a specific group is difficult to find at home but its members are co-operative once they are contacted, the biases would tend to balance each other.

Wilcox (1977) conducted a telephone survey of adult residents of a certain metropolitan area. Up to three calls were made and attempts were made to obtain at least some demographic items from refusers. The final response rate was 84% while 29% of the refusers provided some demographic information. The results indicated that less well-educated respondents were easier to find at home but were more hesitant to co-operate. Wilcox found that bias in education items was actually increased by adding calls because the probability of co-operation increases across education levels whereas the likelihood of being at home, falls. Wilcox (1977:596) admits that a limitation of his study is the fact that the basis of comparison was total response after three calls.

A general conclusion that can be made from this study is that both sources of non-response bias should be considered in the design of corrective techniques and estimation of effects. According to Wilcox (1977:593):

> *The implications of such interaction are significant from several perspectives. Because most methodologies are designed to treat only one source of error, application of a single technique could result in increased error, particularly if the interaction is of an offsetting*

nature. If resources are allocated to correct for not-at-homes, refusal bias might predominate. In contrast, if special effort is devoted to refusals, not-at-home bias may overly influence the results.

The same issue has also been addressed in section 2.3.2.2.

3.6.10. Conclusion

The extensive literature[2] of the non-response problem contains numerous empirical analyses which show that there are important differences between respondents and non-respondents in most surveys. Most of these studies focus on demographic, socio-economic and personality variables. The most established findings are summarised in Table 3.4 but the precise pattern of differences will vary from survey to survey and from country to country. No evidence of how these findings apply to RSA populations could be obtained.

Table 3.4 Differences between respondents and non-respondents

	More likely to refuse	Less likely to refuse	Difficult to contact	Less difficult to contact
Income	Lower income	Higher income	Higher income	Lower income
Education	Lower education level	Higher education level	Higher education level	Lower education level
Occupation	Blue-collar worker	White collar worker	Employed outside home	Employed inside home
Household size			Smaller family	Larger family
Age	Elderly			Elderly
Interest in subject	Less interested	Interested in subject		
Children			No young children in home	Young children in home

Finally, it is important that the estimation of differences between respondents and non-respondents on *auxiliary* variables should not become the sole objective of identification studies. Kanuk and Berenson (1975:449) state:

Even if reliable differences are found to exist between respondents and non-respondents, the problem remains of estimating the effects of those differences on the questions which are the object of the survey.

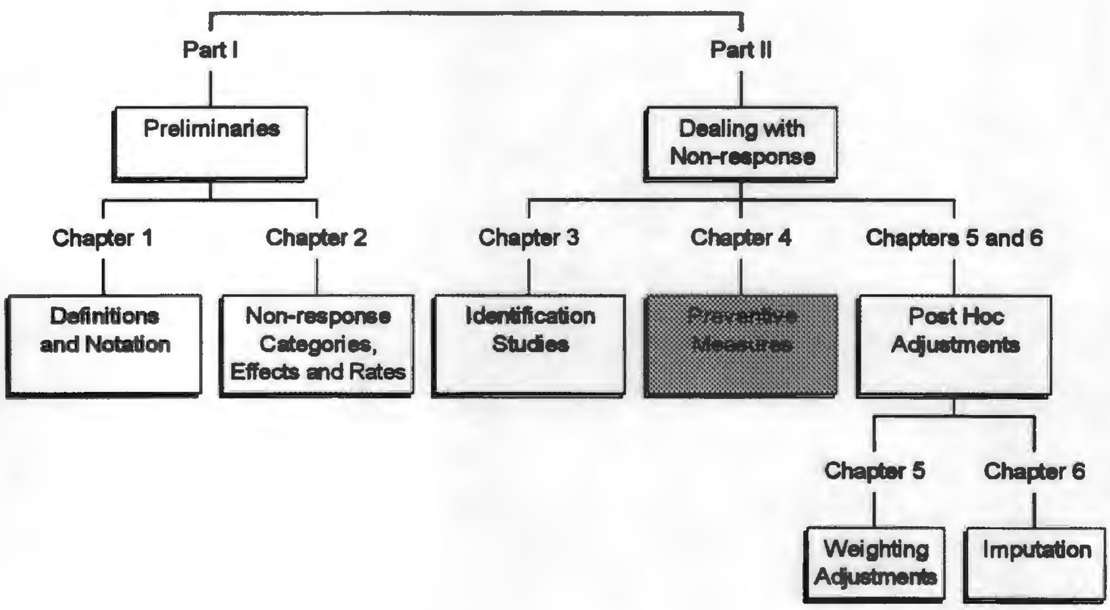
---

[2] See also Groves (1989: Chapter 5) for a review of "the more consistent findings" in literature, as well as Scott (1961) and Kanuk and Berenson (1975) for a review of various studies dealing with differences between respondents and non-respondents. Other literature sources are Pavalko and Lutterman (1973), Suchman (1962), Williams (1968), Wiseman and McDonald (1979), Mandell (1974) and Massey, Barker and Hsiung (1981). See Ferber (1966) for an identification study dealing with bias due to *item* non-response.

CHAPTER 4

DESIGN OF SURVEYS TO MINIMISE NON-RESPONSE:

PREVENTIVE MEASURES



CHAPTER OUTLINE

CHAPTER 4

DESIGN OF SURVEYS TO MINIMISE NON-RESPONSE:

PREVENTIVE MEASURES

"Thanks for having such patience with me and for making me ashamed of myself to a degree where I find myself resolving not to procrastinate in any form or manner whatsoever."

(Note received together with the completed questionnaire after the *sixth* follow-up letter in a mail survey.)
John Goyder, The Silent Minority

4.1. INTRODUCTION

It was shown in Chapters 2 and 3 that non-response *may* cause serious biases in survey estimates through its two components: the non-response rate and the magnitude of the differences between respondents and non-respondents. A number of studies were discussed which showed that there are usually systematic (non-random) differences between respondents and non-respondents in surveys. It was also shown that classical design-based inferences are no longer possible when there is non-response in the survey. However, non-response does not necessarily cause irreparable harm: there exists a large body of knowledge on how to deal with non-response (Moser & Kalton 1971:167). The approaches to dealing with the non-response problem are two-fold: (1) *preventing* the problem from becoming too large, i.e., attempting to reduce or (ideally) eliminate non-response and (2) applying *post hoc* adjustment procedures, usually during the data processing phase and the estimation and analysis phase of the survey.

Preventive measures, the subject of the present chapter, involve "non-statistical steps and data-collection strategies that are taken before or during field operations with the intention of increasing the likelihood that elements in the population, if they were selected for the survey, would participate" (Kalsbeek 1980:134). Preventive measures generally focus on:

1. Motivating the *sample element* to respond
2. Skilful design of the *questionnaire* so as to increase the likelihood of response
3. Increasing *interviewer* ability to obtain satisfactory response rates

Post hoc methods, which consist mainly of *weighting* and *imputation* procedures, are discussed in Chapters 5 and 6.

Preventive measures have variously been classified into *preliminary*, *concurrent* and *follow-up* techniques (Kanuk & Berenson 1975:441). Preliminary techniques are usually applied *prior* to the principal wave of data collection while concurrent techniques are incorporated into the first and/or subsequent data collection waves. Research literature is dominated by numerous tests of preliminary and concurrent techniques, typically manipulated one or two at a time. Examples of such techniques tested include: financial and material incentives, personalisation of correspondence, questionnaire lay-out and length, colour of printed matter, type of outgoing postage, type of return postage, content of cover letter, endorsement of the survey, and many more. A number of preliminary and concurrent techniques are briefly discussed in section 4.2; more extensive discussions can be found in various sources, for example Warwick and Lininger (1975), Moser and Kalton (1973), Bailey (1987), Scott (1961), Dillman (1978, 1983, 1991) and Kanuk and Berenson (1975).

Substitution of non-respondents with respondents in the survey is a technique which is often used to deal with non-response. Substitution is neither a preliminary nor a concurrent technique and is discussed, in a somewhat isolated position, in section 4.3.

Follow-up techniques, i.e., efforts to improve the response rate *subsequent* to the principal wave of data collection, have proven to be more successful than preliminary and concurrent techniques. A more elaborate discussion of follow-up techniques is presented in section 4.4.

In many surveys, follow-ups are restricted to a sub-sample of non-respondents. Sub-sampling of non-respondents is discussed in section 4.5.

4.2. PRELIMINARY AND CONCURRENT TECHNIQUES

The expected response rate of a survey may be greatly influenced at the planning stage of the survey by examining the possible effects that various alternative survey operations may have on response (Särndal *et al.* 1992:564). Three factors that affect the response rate (but over which the surveyor usually has little control) are the *survey organisation* and/or the organisation for whom the survey is being conducted, the *target population* and the *subject matter* of the survey (Moser & Kalton 1971:263). If the survey organisation or the organisation for whom the survey is being conducted lacks credibility in the population, the endorsement of a "respected" organisation, for example, a well-known survey or research organisation, a university or a government department[1] may be obtained. Although there is little a surveyor can do about the population or the subject matter of the survey, he/she may select a *data collection method* which is most suited to it (see Dillman 1978).

4.2.1. Survey Credibility

A number of studies have indicated that the perceived credibility of the survey organisation and/or the organisation for whom the survey is being conducted, may assure sample elements of the survey's legitimacy and value (Kanuk & Berenson 1975:450). "Official" or "respected" organisations, for example, legitimate scientific institutions such as the Human Sciences Research Council (HSRC), governmental departments, universities or well-known non-profit organisations may tend to have higher response rates than organisations that may seem to have ulterior motives, such as commercial organisations or organisations that are not well-known (Bailey 1987:154). This may be especially true in business surveys or in surveys of professional people (Paxson, Dillman & Tarnai 1995:303).

In a randomised experiment using three different letterheads on a mailed questionnaire - that of a government department, a university and a market research firm - Scott (1961) found no significant differences in response rates. However, when the response rate to the governmental survey was compared with the pooled results of the two non-governmental organisations, a significant advantage for the governmental survey was revealed.

In a randomised experiment in Maryland, USA, Brunner and Carroll (1969) tested the effectiveness of a prior letter by mail, one bearing the letterhead of the University of Maryland and the other that of an unknown consulting firm. A significantly higher completion rate was obtained for those who received advance

[1] A government department will not necessarily have credibility among the population, but at least the government may have the image of being able to enforce participation!

notification on the university letterhead (72,5%) as opposed to those who received prior notification on the consulting firm's letterhead (52,3%).

4.2.2. Choice of Data Collection Method

There are various factors that a surveyor should consider when choosing a suitable data collection method, for example, its relative cost, its expected response rate, the characteristics of the population and the subject matter of the survey. While the cost of mail surveys is considerably lower than that of face-to-face interview surveys, the response rates obtained in mail surveys are relatively low - sometimes as low as 10% (Kish 1965:538). Telephone surveys are cheaper than face-to-face interview surveys but their response rates are reported to remain at least 5% below response rates obtained in face-to-face interview surveys (Bailey 1987:202). Although face-to-face interview surveys are the most expensive of the three data collection methods, they tend to generate the highest response rates.

The characteristics of the population should be studied before deciding which data collection method to use. For example, telephone surveys will lead to coverage error if a large proportion of the population does not own a telephone. Furthermore, certain sub-groups in the population, such as the less educated, the elderly or those in the lower occupational categories, have lower response probabilities (see Chapter 3). In surveys where these sub-groups constitute a large proportion of the population, mail questionnaires may not be suitable. On the other hand, the subject matter of the survey is also significant when choosing a suitable data collection method: several studies have shown that the mail survey is superior to the face-to-face interview for gathering information on sensitive or controversial subjects. (Moser & Kalton 1971:257.)

The various data collection methods may be combined to make use of their different strengths (Moser & Kalton 1971:239). Mail questionnaires are often supplemented by follow-up telephone (or face-to-face) interviews to increase the overall response rate (see section 4.4). A disadvantage of using different data collection methods in the same survey is that measurement errors may differ among the methods.

4.2.3. Advance Notification

A number of researchers (e.g., Dillman, Sinclair & Clark 1993; Brunner & Carroll 1967, 1969; Jobber & Sanderson 1981) have tested the effectiveness of using *advance notification* by mail or telephone to increase the response rate. In face-to-face interview surveys, the suggestion is that by contacting the sample elements beforehand and arranging an interview it may be possible to (1) reduce the proportion of non-contacts; (2) enhance the credibility of the interviewer and (3) reduce fear and suspicion among sample elements. At the same time, however, such a procedure may (1) increase the refusal rate since individuals may find it easier to refuse an interviewer over the telephone (or by mail) than in person and (2) some individuals may use this information to make doubly sure of being out or not answering the door at the time of the interview.

In a controlled experiment in Maryland, USA, Brunner and Carroll (1967) found significantly lower completion rates among sample elements who received a prior telephone call arranging an appointment for an interview. This was due to the much greater refusal rate received by interviewers over the telephone as compared to that received by interviewers at the door. However, in a subsequent study (mentioned in section 4.2.1), Brunner and Carroll (1969) found that using advance notification by mail has a positive effect only in some situations: when the letterhead of the University of Maryland was used, advance notification was effective in increasing the response rate; no significant effect was found when using an unknown consulting firm's letterhead.

4.2.4. Postage and Return Envelopes

The inclusion of either a stamped return envelope or a business reply envelope seems to be a generally accepted practice in mail surveys (Kanuk & Berenson 1975:443). Bailey (1987:160) mentions a study (conducted in 1951) in which a stamped, self-addressed envelope was compared with no envelope at all. A response rate of 66% was obtained for the envelope and only 12% for no envelope.

Bailey (1987) and Kanuk and Berenson (1975) discuss a number of studies that have attempted to determine the effect of type of postage on both the outgoing and return envelopes in mail surveys, for example, air or registered mail versus regular mail. Somewhat inconsistent results have been obtained. *In general*, a greater response rate is expected from hand-stamped return envelopes than from business-reply envelopes (Bailey 1987:161). The reasoning behind this is that a stamp may attract more attention and that a person may feel guilty if he/she throws it away. Another reason (especially in rural areas) is that some sample elements may not know that postage will be paid on a business-reply envelope. On the other hand, a hand-addressed envelope with a stamp may seem unprofessional while a printed envelope from someone with franking privileges may seem "authoritative, legitimate, formal and prestigious" (Bailey 1987:161). The target population should therefore be taken into consideration when deciding on the type of postage to use.

In a randomised experiment, Kernan (1971) found that neither personalised addressing nor the use of a hand-affixed postage stamp on the *outgoing* mail had a significant effect on response rates. In a randomised experiment by Gullahorn and Gullahorn (1963), a stamped *return* envelope was found to have a significantly higher response rate than a business-reply envelope.

4.2.5. Personalisation of the Cover Letter

In mail surveys, a covering letter is usually sent out with the questionnaire explaining why and by whom the survey is undertaken, how the sample element has come to be selected for the survey and why he/she should take the trouble to reply (Moser & Kalton 1971:264). The effects of personalising the cover letter have been explored by many researchers, but the results are inconclusive (Kanuk & Berenson 1975:444).

Scott (1961) discusses a number of studies dealing with the style of the cover letter, the salutation (for example, "Dear Mr. Smith", "Dear Madam" or "Dear Occupant") and the use of hand-written addresses and signatures versus facsimiles. No significant evidence was obtained that personalisation increases the response rate. (See also Kernan (1971) for a similar result.)

In mail surveys, personalisation of the cover letter obviously decreases anonymity (see section 4.2.9) unless the cover letter is convincingly detached from the (unmarked) questionnaire.

4.2.6. Interviewer Introduction

In face-to-face or telephone interview surveys, the image of the survey organisation as portrayed by the interviewer may be an important factor in respondent motivation. The aim of the introductory procedures must be to increase the sample element's motivation to co-operate. The interviewer will usually begin by stating the organisation he/she represents and giving a brief statement of why the survey is being conducted. The interviewer may need to explain precisely why and for whom the survey is being conducted, what is expected to emerge from it, to whom the results will be of interest, and so on. (Moser & Kalton 1971:274.) In face-to-face interview surveys, interviewers must present official identification and the survey material must convey an appearance of being "official" (Platek 1977:199).

4.2.7. Questionnaire Length

Common sense suggests that shorter questionnaires should result in higher response rates than longer ones, however, questionnaire length has been a controversial issue in survey research (Kanuk & Berenson 1975:442). Adams and Gale (1982:232) contribute discrepancies in conclusions reached in past studies to methodological problems. In a controlled experiment, i.e., holding other influences on the response rate to a minimum, Adams and Gale (1982) found significant differences in response rates for a 1-page, 3-page and 5-page questionnaire. The 3-page questionnaire had a significantly higher response rate than the 1-page questionnaire, while the 5-page questionnaire had a significantly lower response rate than both the 1-page and the 3-page questionnaires. The 1-page questionnaire is therefore not the most desirable length for a high response rate. However, the 1-page questionnaire had a response rate nearly twice that of the 5-page questionnaire.

According to Dillman (1993:302) attention should be given to the combined effect of "respondent-friendly design" and questionnaire length. Poorly constructed questionnaires - those that have no instructions or inadequate instructions for completing questions, those that have unclear response categories or have too many open-ended questions, or questionnaires asking difficult questions - may result in poor response, leading to non-response bias and measurement errors (Bailey 1987:156). It is a good idea to have the questionnaire pre-tested in order to improve its clarity and acceptability (Daniel 1975:294).

Childers and Ferrell (1979) suggest that the perceived length of the questionnaire may be affected by a wide variety of factors, including number of questions, number of pages and the physical size of the pages.

4.2.8. Questionnaire Size and Colour

In a randomised experiment by Childers and Ferrell (1979), a significantly higher response rate was obtained for an $8\frac{1}{2}\times 11''$ sheet of paper for the questionnaire than a legal-size $8\frac{1}{2}\times 14''$ sheet. The number of pages in the questionnaire did not have a significant effect on completion rates, although the direction of the difference in response rates was in favour of the smaller number of pages (one sheet).

The use of coloured paper has been explored by some researchers. Gullahorn and Gullahorn (1963) found no significant difference in response rates for questionnaires printed on green paper versus questionnaires printed on white paper. (See also Jobber and Sanderson (1981) for a similar result in a business survey.)

4.2.9. Anonymity and Randomised Response Techniques

It has generally been assumed that assurances of anonymity encourage a high level of voluntary response and that assurances of anonymity minimise measurement error when response is mandatory (Kanuk & Berenson 1975:446). However, experimental evidence indicates that the promise of anonymity to sample elements in mail surveys - either explicit or implied - has no significant effect on response rates. On the other hand, the degree to which sample elements welcome anonymity would seem to be largely dependent upon the nature of the study. In studies of more controversial or sensitive subjects, failure to ensure anonymity may increase the non-response rate.

Ensuring true anonymity in mail surveys is not always easy, especially if follow-ups are to be conducted. A common practice in mail surveys is to identify respondents by a serial number printed on the questionnaire. In this case, the sample element cannot be promised *anonymity* but only *confidentiality*. Various devices have been reported for preserving anonymity without sacrificing the knowledge of which addressees have responded[2]. The use of invisible ink to blind-code questionnaires is not unusual, although this practice cannot be considered an ethical way of identifying respondents. Other studies have been reported in which sample elements were asked to return a separate postcard bearing their name and address to indicate that they had mailed the questionnaire under separate cover. (Kanuk & Berenson 1975:446.) Another (less ideal) solution is to send follow-up letters to all sample elements with an appropriate sentence to "ignore this letter if you have already returned the questionnaire".

In telephone surveys, a potential respondent may be more difficult to convince that his or her response is anonymous, while in face-to-face interview surveys the technique of *randomised response* may be used to increase the response rate.

The ingenious technique of randomised response, proposed by Warner (1965), was designed for face-to-face interview surveys dealing with sensitive issues. It was designed with the purpose of reducing the level of self-disclosure required of a respondent and, hopefully, to lead him or her to give truthful responses (i.e., to reduce measurement errors). The basic idea of the technique is that each element in the sample chooses to answer one of two (or more) questions, one of which is the sensitive question. The choice of which question to answer is made by means of a random device provided by the surveyor (for example, an ordinary die). The surveyor knows the selection probability for each question. The respondent employs the random device to choose one of the questions and answers "yes" or "no" without disclosing to the interviewer which question he/she is answering. (Moser & Kalton 1971:328.) Knowledge of the question selection probabilities allows unbiased estimates of the distribution of answers to the sensitive question to be calculated without knowing which question a specific respondent has answered. However, a disadvantage of the technique is that the variance of estimators is increased.

The effectiveness of the randomised response technique is related amongst others, to the specific item of interest, the choice of alternative items, whether the

---

[2] A knowledge of which addressees have responded is important in order to avoid having to send follow-up letters to all elements in the sample.

respondents believe in the promised confidentiality and the privacy with which the randomisation procedure is performed (Madow, Nisselson & Olkin 1983a:76).

Särndal *et al.* (1992:572) list a number of recent references on the randomised response technique.

4.2.10. Financial and Material Incentives

A number of studies (Gunn & Rhodes 1981; Chromy & Horvitz 1978; Ferber & Sudman 1974; Kerachsky & Mallar 1981; James & Bolstein 1990) point to improved response rates in most surveys if an incentive of some kind is used. In a controlled experiment, Dohrenwend (1970-71) reported no significant improvement in response rates if a honorarium of $5 is offered.

Not only cash payments but also material incentives such as stamps, lottery tickets, miniature pencils and ball-point pens and entries into lucky draws have been used to (hopefully) stimulate response. Factors that affect the potential success of incentives include the amount of compensation offered, the required period of co-operation (e.g., single interview, record keeping, panel participation), the auspices under which the survey is conducted, the socio-economic status of the sample elements and the nature of the information sought (Chromy & Horvitz 1878:473).

An important consideration is the effect (if any) that incentives have on measurement errors in a survey (Bailey 1987:158). The ethics of using monetary incentives to "enforce" responses to a survey should also be considered.

4.2.11. Deadline Date

In some studies it was found that a deadline date led to a higher response rate in the *initial* returns (before a follow-up message was sent), however, a deadline seemed to make little difference in the final overall response rate (Bailey 1987:158). The advantage of a deadline is that it may prevent the sample element from putting off the completion of the questionnaire. On the other hand, a deadline which is set too far in the future may cause those who would otherwise have replied immediately, to wait because "there is no rush". Alternatively, if the deadline date has already expired, the sample element may discard the questionnaire. According to Bailey (1987:159) it is safer not to use deadlines, particularly if follow-up letters are to be used.

4.2.12. Clear Interview Assignment Materials

In face-to-face interview surveys, the core of the interviewer's task is to locate the sample elements, to obtain interviews with them and ask the questions and record the answers as instructed (Moser & Kalton 1971:273). Sufficient information to locate sample elements, i.e., correctly spelt full name, complete street address with house number or description, home or business telephone number and name and address of current employer are important measures to reduce the proportion of non-contacts. Clear interview assignment materials may also reduce the potential for measurement errors and item non-response. In mail surveys, complete and up-to-date address lists are imperative while in telephone surveys, up-to-date directories should be used unless random digit dialling is used.

In surveys where the information on the sampling frame is incomplete or outdated, intensive tracing efforts are required. Tracing involves pursuing leads based on information available from the sampling frame, directory assistance, informants such as other members of the household or neighbourhood, publicly available records or previously completed questionnaires. (Lessler & Kalsbeek 1992:168.) Interviewer recruitment and training (see section 4.2.13) should take into account the skills of prospective interviewers to find addresses in the field, read maps and trace sample elements who have moved.

4.2.13. Recruitment, Training and Supervision of Interviewers

The success of a face-to-face or telephone interview survey is very much dependent on the way the interviewer presents the survey to the sample elements (Platek 1977:202). Moser and Kalton (1971:285) describe some desirable personal characteristics of interviewers; namely honesty and integrity, interest in the work, accuracy, adaptability, pleasant and business-like manner and sufficient intelligence to understand and follow complicated instructions. Female interviewers are expected to achieve higher response rates in some surveys than male interviewers, because women tend to arouse less suspicion and pose less of a threat to certain segments of the population. However, there is little experimental evidence to suggest a relationship between response rates and the gender of the interviewer. (Lessler & Kalsbeek 1992:169.)

Supervision of interviewers in face-to-face and telephone surveys is essential, both to detect bad work and to keep interviewers up to the mark. Faked interviews (often referred to as "curbstoning") refers to a situation in which the interview does not take place but the interviewer fills in the questionnaire or some missing information on a partially completed questionnaire, and pretends that the interview occurred or that the responses were complete (Chapman 1983:45). To help reduce the frequency of curbstoning, fieldwork checks should be done on a large proportion of questionnaires to verify that the interviews have occurred (Moser & Kalton 1971:292). Telephone calls to a number of "respondents" or "non-respondents" may be made to verify whether an interviewer has in fact called, or whether the sample element has in fact refused or been away from home. The interviewers should be informed that validation will occur soon after the interviews have been conducted. The validation process should be convincing to the interviewers.

According to Warwick and Lininger (1975:187) interviewers in less developed countries often find that sample elements do not understand what is meant by "research" and "personal interviews". For this reason, there is often strong suspicion of any organised data-gathering activity, including the most thoroughly legitimated national census. This may lead to high refusal rates in such countries or areas. In general, the successful conversion of refusals into responses depends very much on the interviewer's experience and his ability to find arguments to refute the reasons for refusals[3] (Lessler & Kalsbeek 1992:169). Durbin and Stuart (1954) found the refusal rate for inexperienced amateur interviewers to be about three times that of experienced professional interviewers. In most instances, a hard-sell strategy at the initial refusal may be detrimental to the response rate: conversion attempts are most effectively made a few days after the initial refusal (Lessler & Kalsbeek 1992:170).

It is the task of supervisors to ensure that interviewers remain convinced of the importance of the survey and the validity of the arguments they deliver to persuade sample elements to respond. Remuneration, working hours, working conditions and work-load of interviewers affect their morale and efficiency. Other contributing factors

---

[3] Of course, in mail surveys the reasons for refusals cannot be determined unless a telephone or face-to-face interview is used.

to low morale are difficult travel conditions, frustration in locating sample elements and assignments in areas with high crime rates. (Moser & Kalton 1971:295.)

4.2.14. The Total Design Method

The Total Design Method (TDM) can be described as a system of procedures and techniques that attempt to maximise response rates in ways consistent with obtaining "quality" responses. The emphasis is not on a particular technique (e.g., personalisation or follow-ups) but on how these procedures can be linked to influence questionnaire recipients positively. (Dillman 1991:233.) The theoretical framework used in this approach proposes that questionnaire recipients are most likely to respond if they expect that the perceived benefits of doing so will outweigh the perceived costs of responding. Thus, every visible aspect of the survey is subjected to three design considerations: (1) the reduction of perceived costs (e.g., making the questionnaire appear easier and less time-consuming to complete), (2) increasing perceived rewards (e.g., making the questionnaire itself interesting to fill out by adding interest-provoking questions) and (3) increasing trust that the promised rewards will be realised (e.g., by using official letterheads and endorsements). (Dillman 1991:234.)

The TDM includes recommendations on the ordering of questions to ensure that interesting ones related to the topic described in the cover letter come first; individually printed, addressed and signed letters; the use of smaller than usual business stationery to reduce costs and make the entire request appear smaller and easier to comply with; cover letter content that includes descriptions of the study's social usefulness and why the recipient is important (repeated in different ways in each of the subsequent mailings); an explanation of identification numbers and how confidentiality is protected. The TDM also recommends the use of four carefully spaced mailings, including a post-card follow-up one week after the original mailing, a replacement questionnaire and cover letter informing the recipient the questionnaire has not yet been received four weeks after the initial mailing, and a second replacement questionnaire and cover letter seven weeks after the first mailing, sent to non-respondents by certified mail.

Other details of the TDM, in particular how they are integrated to create a holistic effect, are described in Dillman (1978, 1983). Dillman (1991:234) states:

> *The major strength of the TDM as a comprehensive system is that meticulously following the prescribed procedure consistently produces high response rates for virtually all survey populations. Response rates typically reach 50-70% for general public surveys and 60-80% for more homogeneous groups where low education is not a characteristic of the population .... Response rates greater than 80% have frequently been reported .... To my knowledge, no study that has utilised a 12-page or smaller booklet and followed the TDM in complete detail, from questionnaire through the full set of implementation procedures, has obtained less than a 50% response rate.*

A significant conclusion that can be made regarding preliminary and concurrent techniques, is that there is no single "magic bullet" to secure high response. Reasonably high response rates can, however, be obtained through the use and integration of multiple techniques, including *follow-ups* (see section 4.4). Despite the large amount of research on preliminary and concurrent techniques to stimulate the response rate, the follow-up is the only technique which has consistently been found to raise response by a *substantial* amount.

4.3. SUBSTITUTION

Substitution, the replacement of non-respondents with (presumably similar) population elements not originally selected for the sample, is a procedure that is often used to "deal" with unit non-response. In general, two basic types of substitution procedures are used (Chapman 1983:45):

1. Selection of a random substitute
2. Selection of a specially designated substitute

4.3.1.Random Substitution

With a random substitution procedure, one or more "back-up" population elements are selected on a probability basis to replace each non-respondent. (More than one back-up is often selected for each non-respondent to allow for non-responding substitutes.) The use of random substitution has the advantage that inclusion probabilities of substitute elements are known. If substitutes cannot be obtained for some non-respondents, other adjustment procedures, such as weighting or imputation (Chapters 5 and 6) are used. If the population can be divided into homogeneous sub-groups, the substitute for a particular non-respondent is usually chosen from the same sub-group of the population, e.g., the same block, area or stratum as the non-respondent.

For many random substitution procedures, back-up elements are selected *prior* to the data collection phase of the survey. This avoids any delay and trouble that would be involved in selecting a random substitute in the field, e.g., having to train interviewers how to select a *random* replacement from the same population sub-group as the non-respondent.

According to Chapman (1983:48) the use of substitution is most appropriate for a survey that involves a deeply stratified, relatively small sample, for example, surveys of institutions such as schools or hospitals in which a substantial amount of stratification information is available. Chapman (1983) discusses a random substitution procedure used in a study reported by Williams and Folsom (1977). In the first stage of the survey, a probability sample of four schools is selected from each of 600 strata: two of the four schools are randomly selected to be used for the initial sample, while the other two are designated as back-ups. If either one or both of the schools selected for the initial sample do not respond, one or both back-up schools are approached, as needed, to be substitutes. If the substitutes also decline to participate in the survey, no other substitutions are used and instead, weighting adjustments are made for the non-responding schools.

4.3.2. Over-sampling and Supplementary Sampling

According to Deming (1953:34): "It is important to bear in mind that the problem of *non-response is not solved by starting off with an excess of cases to allow for shrinkage.* There is no substitute for response". However, *over-sampling*, i.e. selecting more than the required number of elements to compensate for non-response, is often used to increase the number of responses obtained. The number of elements to select is obtained by multiplying the required sample size by the inverse of the expected response rate.

A set of smaller *supplementary* samples is sometimes selected in addition to the initial sample and using the same sampling design as the initial sample, in case non-response in the initial sample is higher than expected (Lessler & Kalsbeek 1992:176). Suppose the response rate in the survey is expected to be between r_L and r_H. An initial sample of size $\frac{n}{r_H}$ is selected, as well as a set of m independent supplementary samples of size $\frac{1}{m}\left(\frac{n}{r_L}-\frac{n}{r_H}\right)$. As the survey progresses, individual supplementary samples are added until a sample size is obtained which is as close to n as desired. A full commitment must be made to obtaining responses from each element as each supplementary sample is added. Clearly, the larger m is allowed to be, the closer one can get to the desired sample size and the more supplementary sampling resembles individual substitution. This method is often used in one-time surveys in which there is relatively little prior knowledge of response rates in the population or where making individual substitutions is too difficult or costly. (Lessler & Kalsbeek 1992:176.)

It should be noted that, when using over-sampling or supplementary sampling, response rates are still calculated based on the actual sample size; i.e., if a sample of size 100 is desired but a sample of size 200 is selected, then, if 100 elements respond, the response rate is 50% and not 100%.

4.3.3. Deterministic Substitution

Substitutes may be identified not by probability sampling, but by applying (often during data collection) a pre-determined set of criteria for one or more back-up elements. For example, in an area household sample, the fieldworker may be instructed to approach the dwelling unit immediately on the right of the non-responding unit or, more generally, the next encountered population element with certain characteristics similar to those of the non-respondent.

Chapman (1983) discusses a deterministic substitution procedure used by Sirken (1975) in a multi-stage household sample of 2100 households. Interviewers were instructed to make up to three calls at a household. If these attempts failed, the household was to be dropped from the sample. The interviewer was then to approach the household directly to the right of the non-responding household. If this attempt also failed, the household to the left of the non-respondent was to be approached. Only one call was made to each substitute household. If the substitutes failed to respond, other households were approached in some manner (not described by the author) until a substitute household was obtained. Approximately one-third of responses in the survey was obtained from substitutes.

Sirken (1975) compared the values of four survey variables across three age categories for three groups of respondents:

1. Initially selected individuals who responded on the first call
2. Initially selected individuals who responded on the second or third call
3. Substitute individuals

The comparisons indicated that the characteristics of the substitute individuals were very much the same as the initially selected individuals who responded on the first call. For the oldest age group (35 years and older), the characteristics of the substitutes were also approximately the same as those of persons interviewed on the second or third call. However, for the two age groups "13 to 17 years of age" and "18 to 34 years of age" differences were found between the first-call respondents and substitute respondents compared with persons interviewed on the second and third call. Sirken speculated that the characteristics of non-respondents might be closer to those of persons interviewed on the second or third call than to those of the substitutes. However, since the characteristics of non-respondents were not known in this survey, these remain speculative results.

4.3.4. Advantages and Disadvantages of Substitution

There are two disadvantages that generally apply to substitution procedures (Chapman 1983:49):

1. A back-up element may be seen as one that is just as good (or nearly as good) as the element originally selected. As a consequence, the effort extended to obtain responses from the original sample may not be as intense as it would be if no substitutes were available. This may lead to a higher non-response rate which may produce greater biases in the survey results.

2. The level of substitution used may easily be ignored when the survey response rate is calculated, i.e., substitute respondents may be viewed as if they were elements selected in the original sample. The survey response rate will be over-estimated and the potential for non-response bias will be under-estimated.

There are two advantages that apply to substitution procedures (Chapman 1983:49):

1. It is possible to reach the targeted number of responses, n_r

2. The number of responses obtained (n_r) is greater than what it would have been if no substitutions were made, hence, the variances of survey estimates are reduced

Obtaining the targeted number of responses has certain practical advantages: (a) if the survey employs a self-weighting design, the final sample will still be essentially self-weighting (Chapman 1983:50) and (b) variance estimation problems caused by strata with fewer than two sample elements can be avoided. It should be noted that, although the final sample (including substitutes) is often treated as "essentially self-weighting", the inclusion probabilities of the substitutes differ from the inclusion probabilities of the elements that they are replacing. Furthermore, their inclusion probabilities are unknown, since they depend on the unknown response probabilities of the elements in the initial sample.

Whenever substitutes are used in a survey, care should be taken (1) to ensure that the maximum effort is made to obtain responses from the original sample elements, (2) to supervise fieldwork, including validating a substantial proportion of the substituted elements to verify that substitutes were needed, (3) to keep accurate records of which elements are substitutes, (4) to report the level of substitution, and (5) to treat the substitutes as "non-respondents" when calculating the survey response rate. (Chapman 1983:50.)

4.3.5. The Impact of Substitution on Survey Estimates

Substitution procedures are often criticised for not being effective in reducing non-response bias. The main criticism is that the non-respondents in the survey differ from the substitutions simply because the latter responded while the former did not. Substitutes therefore resemble the responses that are already in the sample rather than the non-responses (Kish 1965:549). According to Chapman (1983:48):

> *This is an unfair criticism if directed solely at the use of substitution procedures, since all the methods used for non-response imputation, including weight-adjustment procedures, suffer from that same basic weakness: data for non-respondents have to be supplied (imputed) from data provided by respondents.*

The technique of random substitution within sub-groups of the population resembles sample or population weighting adjustments for non-response and is also conceptually similar to randomised hot-deck imputation used to deal with item non-response (see Chapters 5 and 6). The random substitutes have the same expected values as the respondents in the sub-groups from which the substitutes are selected - although not necessarily the expected values of the non-respondents. For example, if a simple random sample is used to select the original sample and the substitutes within a sub-group, the bias of the resulting estimate will be the same as that of an estimator obtained from a weighting adjustment or hot-deck imputation procedure - assuming that the sub-groups are defined in the same way as the weighting or imputation classes. (Lessler & Kalsbeek 1992:175.)

Substitution procedures will be effective in reducing non-response bias to the degree that each substitute resembles the sample element it is replacing. If substitutes and non-respondents can be matched on the basis of known auxiliary variables correlated with the survey variable, the biasing effect of non-response will be diminished for the same reasons that weighting class adjustments and hot-deck imputation reduce bias. Unfortunately, as with most remedies for non-response, substitution does not completely eliminate bias. (Lessler & Kalsbeek 1992:176.)

According to Chapman (1983:49), if there is relatively little auxiliary information available about non-respondents, substitution would probably not provide any *improvement* in terms of bias reduction over the use of weighting adjustments. On the other hand, some reduction in survey variances would result due to an increase in sample size, but the small reduction in survey variances may not be worthwhile considering the potential disadvantages of substitution procedures.

No theoretical results are available for the *deterministic* substitution procedure. Such a model would have to be complex to appropriately reflect the relation between the characteristics of the substitutes and those of the non-respondents or between those of the substitutes and those of the respondents. In general, because of the usual physical proximity between the substitute and the non-respondent, the deterministic substitution procedure may be an improvement over a technique such as mean value imputation (see sections 6.2.3 and 6.2.4).

4.4. CALL-BACKS AND FOLLOW-UPS

The most successful method for reducing the survey non-response rate is *calling back* repeatedly on non-respondents or sending *follow-up* letters to urge those who have not yet returned their questionnaires, to do so. Scott (1961:154) called the use of follow-ups in mail surveys "the most potent technique yet discovered for increasing the response rate". In mail surveys, follow-up letters are sometimes used in *combination* with a telephone call. Two or three reminder letters are mailed (sometimes accompanied by another copy of the questionnaire) to obtain as many responses as possible from the mail phase before the more expensive telephone follow-ups are conducted (Särndal *et al.* 1992:565).

In mail surveys, the use of call-backs usually requires that non-respondents are identifiable - if not, follow-up questionnaires will have to be sent to the entire sample which, in addition to being a more costly option, may add to the chagrin of refusers or reluctant respondents. Furthermore, some respondents who have already returned their completed questionnaires may think theirs have not been received by the survey organisation and also complete and return the second questionnaires.

4.4.1. Cost of Call-backs in Face-to-face Interview Surveys

In face-to-face interview surveys, call-backs are generally expected to be much more expensive *per response* than the initial calls because of the increased field costs in later calls. Field costs often increase in later calls because of increased travel costs and the additional effort required to locate the non-contacts (Cochran 1977:366). However, the cost increase of later calls relative to first calls is often over-estimated. According to Durbin (1954:74), if the success rates of successive waves of calls increase (e.g., as is initially the case in column 2 of Table 4.1), it may actually be less expensive in terms of *cost per response* to continue re-calling than to confine the interviewing to the initial calls. An excessive cost increase may only be experienced as the interviewer nears the end of his/her assignment (as in column 3 of Table 4.1), because many *second* calls will be made at a comparatively early stage of the field work, so that it should be possible for the interviewer to fit them into his/her route in an economical manner. Furthermore, according to Kish (1965:552) the higher cost per response in fourth, fifth and later calls may have only modest effects on the *cumulative* cost per response because of the small proportions of interviews in later calls (see column 4 of Table 4.1).

It is quite conceivable for the cost per response on a second call to be cheaper than the cost per response on a first call because (1) at the second call the interviewer will probably already have located the address and found the best way to get to it and (2) the interviewer has possibly gained some knowledge of the sample element's movements at the first call.

Durbin (1954:74) illustrates the above reasoning with a practical example of a survey of which the results are summarised in Table 4.1. (The cost figures have been standardised to make the cost of a first call interview unity.)

Table 4.1 ***Response rates and relative costs on successive waves of calls in a selected survey***

	(1)	(2)	(3)	(4)
Call number	Cumulative response rate	Success rate per call	Relative cost per response	Cumulative relative cost per response [4]
1	34,3	34,3	1,00	1,00
2	66,4	48,8	0,84	0,92
3	80,0	40,5	1,12	0,96
4	84,0	19,9	1,74	1,00
5	85,8	11,4	1,90	1,01
>5	86,8	5,5	2,41	1,02

Source: Adapted from Durbin, J. 1954. Non-response and Call-backs in Surveys. *Bulletin of the International Statistical Institute*, 34:76.

Three things are immediately noticeable from Table 4.1 (Durbin 1954:76):

1. The substantial drop to 0,84 for the cost of a successful second call

2. The fact that in this particular situation, the cumulative relative cost per response remains more or less constant for up to four calls

3. The increase in cumulative relative cost per response seems to be negligible even if the interviews are carried on up to the fifth or sixth call

[4] Calculated as the total cost of interviews obtained up to the *k*-th wave divided by the total number of interviews obtained up to the *k*-th wave.

As Durbin (1954:76) points out, the figures in Table 4.1 were obtained in a single, specially selected survey and should therefore be regarded with reserve until they are supported by a sufficient body of evidence.

4.4.2. Deming's Model of the Optimum Number of Call-backs

Although it may be possible to eliminate non-response entirely (at least the not-at-home category) by calling back indefinitely, usually only a limited number of call-backs are possible in practical applications. One reason is that the benefit of the additional information in terms of the reduction in the *MSE* of the survey estimates, will often be outweighed by the cost of further call-backs. Deming (1953) developed a model which allows factors such as the demographic and socio-economic characteristics of the population and the costs of the fieldwork to be incorporated in determining the optimum number of call-backs to be made.

4.4.2.1. Notation and Assumptions

Assume that the population of size N consists of H hypothetical "response classes" defined on the basis of, e.g., the ability and willingness of population elements to respond, their interest in the subject matter of the survey or their probability to be found at home. The number of elements in the h-th response class is denoted as N_h and the proportion of the population in the h-th response class as $W_h = \frac{N_h}{N}$ $(h = 1, \ldots, H)$. The population mean and variance in the h-th response class are denoted as $\overline{Y}_h$ and S_h^2 respectively. The population mean can be written as a weighted mean:

$$\overline{Y} = \sum_{h=1}^{H} W_h \overline{Y}_h \tag{4.1}$$

The probability that an element in the h-th response class responds on or before the k-th call is denoted as φ_{hk}. It is assumed that $\varphi_{hk} > 0$, although the method can be adapted to include the hard-core non-respondents ($\varphi_{hk} = 0$) (Cochran 1977:367).

In a survey involving k calls, let $n_{r_{hk}}$ denote the number of respondents from the h-th response class. The expected number of respondents from the h-th response class is (Cochran 1977:368):

$$E(n_{r_{hk}} \mid n_{r_k}) = \frac{n_{r_k} \varphi_{hk} W_h}{\sum_{h=1}^{H} \varphi_{hk} W_h} \tag{4.2}$$

The total number of responses obtained in the course of the k calls is $n_{r_k} = \sum_{h=1}^{H} n_{r_{hk}}$. The expected total number of responses in the course of k calls is:

$$E(n_{r_k}) = n_0 \sum_{h=1}^{H} \varphi_{hk} W_h \tag{4.3}$$

where n_0 is the initial size of the sample.

Let $\bar{y}_{r_{hk}}$ denote the mean of the $n_{r_{hk}}$ respondents from the h-th response class. It is assumed that:

$$E(\bar{y}_{r_{hk}}) = \bar{Y}_h \tag{4.4}$$

i.e., respondents and non-respondents in the h-th response class have the same mean $\bar{Y}_h$. This assumption will be satisfied when the response classes are homogeneous with respect to the survey variable (see Chapter 5).

The overall sample mean obtained in a survey involving k calls is denoted as $\bar{y}_{r_k}$, where:

$$\bar{y}_{r_k} = \frac{1}{n_{r_k}} \sum_{h=1}^{H} n_{r_{hk}} \bar{y}_{r_{hk}} \tag{4.5}$$

The values of $n_{r_{hk}}$ and $\bar{y}_{r_{hk}}$ are not known since the different response classes are not identified.

4.4.2.2.Optimum Number of Calls-backs

For a fixed n_{r_k} it follows from (4.2) that (Cochran 1977:368):

$$E\left(\bar{y}_{r_k} \mid n_{r_k}\right) = \frac{\sum_{h=1}^{H} \varphi_{hk} W_h \bar{Y}_h}{\sum_{h=1}^{H} \varphi_{hk} W_h} = \bar{Y}_k \tag{4.6}$$

where $\overline{Y}_k$ denotes the mean for population elements that would respond in the course of k calls. Since (4.6) does not depend on n_{r_k}, the (unconditional) bias of the estimator $\bar{y}_{r_k}$ is:

$$Bias(\bar{y}_{r_k}) = \overline{Y}_k - \overline{Y} \tag{4.7}$$

The conditional variance of $\bar{y}_{r_k}$ for a given n_{r_k} is:

$$V(\bar{y}_{r_k} \mid n_{r_k}) = \frac{\sum_{h=1}^{H} \varphi_{hk} W_h \left[S_h^2 + \left(\overline{Y}_h - \overline{Y}_k \right)^2 \right]}{n_{r_k} \sum_{h=1}^{H} \varphi_{hk} W_h} \tag{4.8}$$

Thus for a call-back policy requiring k calls, the conditional MSE of $\bar{y}_{r_k}$ is:

$$MSE(\bar{y}_{r_k} \mid n_{r_k}) = V(\bar{y}_{r_k} \mid n_{r_k}) + (\overline{Y}_k - \overline{Y})^2 \tag{4.9}$$

When the survey variable y is quantitative and binomially distributed, the MSE takes the form (Rao 1983a:42):

$$MSE(\bar{y}_{r_k} \mid n_{r_k}) = \frac{\overline{Y}_k (1 - \overline{Y}_k)}{n_{r_k}} + (\overline{Y}_k - \overline{Y})^2 \tag{4.10}$$

The approximate unconditional MSE is obtained by replacing n_{r_k} in (4.9) or (4.10) by its expected value (4.3).

In a specific survey, the optimum value of k can be calculated for *presumed* values of W_h, $\overline{Y}_h$, S_h^2 and φ_h. Reasonable estimates of these values may be obtained from analyses of call-back data in past surveys with the same types of questions and the same populations. For example, it is usually found that better educated, higher income and more interested population elements are more likely to respond (see Chapter 3). In any survey, the response probabilities may also be defined to depend on factors such as interviewing strategies (for example hours of interviewing) or information to be obtained from neighbours. Note that the *costs* per completed interview at successive calls may also be predicted from repetitive surveys of the same population for the same items.

4.4.2.3. An Example of the Application of Deming's Model

Suppose an interview survey of heads of households is undertaken in a certain city to gather information on entrepreneurial activity in the city. Three response classes are considered adequate to represent the various response probabilities in the survey. The first class is seen to consist of heads of households with higher education levels who are relatively easy to reach after hours and hence, have a relatively higher likelihood to respond. The second class is seen to consist of household heads who are, e.g., blue-collar workers, shift-workers, working mothers who may be attending to their families at the time of the call and hence, may be somewhat reluctant to respond. The third class is assumed to consist of salespeople, businessmen and women and others hard to contact, those unwilling to respond (for example, because they are involved in illegal entrepreneurial activities) or the illiterate. From past experience, the proportions of the population in the city in these three classes are estimated as 0,7, 0,25 and 0,05.

At the first call, the probabilities of obtaining an interview in the three classes (φ_{h1}) are estimated as 0,6, 0,3 and 0,1 for h = 1, 2 and 3 respectively. To estimate the response probabilities in subsequent calls, the values of φ_{hk} for $k \geq 2$ are chosen to be of the form $\left[\varphi_{h1} + (1 - \varphi_{h1})(1 - \alpha_h^{\ k-1})\right]$ where α_h ranges from 0 to 1 (see Cochran 1977:369). The value of $(1 - \alpha_h^{\ k-1})$ increases as k increases to 2, 3, ... to take account of the intelligent enquiry of the interviewer and the co-operation of the respondent. (How the values α_h are to be estimated, is not clear from Cochran (1977).)

The presumed values for W_h and φ_{hk} in this example, are summarised in Table 4.2 and Table 4.3:

Table 4.2 Estimated values of W_h

Response Class	1	2	3
Proportion (W_h)	0,7	0,25	0,05

Table 4.3 Estimated values of φ_{hk}

	Response Probabilities (φ_{hk})		
α_h	0,1	0,5	0,8
Call Number	Class 1	Class 2	Class 3
1	0,6	0,3	0,1
2	0,96	0,65	0,28
3	0,996	0,825	0,424
4	0,9996	0,913	0,539
5	0,99996	0,957	0,631

The values for α_h used to calculate the entries in Table 4.3 are those suggested by Cochran (1977:369).

Suppose the variable of interest is a percentage suspected to be close to 50% (i.e., the proportion of households who own a business or who would consider opening their own business). Using the results from a number of surveys, Rao (1983a:41) found the value of the approximate bias for binomially distributed variables to be between 0 and 0,05, although in some instances it was found to be as high as 0,20.

Table 4.4 shows for $k = 5$ calls:

1. The estimated (cumulative) response rate over k calls
2. The estimated relative cost per interview (based on experience or historical data)
3. The expected total number of interviews that can be obtained for the same cost as an initial sample (one-call only) of size $n_0 = 3000$
4. The estimated bias in $\bar{y}$

Table 4.4 Relative costs per interview, number of interviews and bias

	(1)	(2)	(3)	(4)
Number of calls required k	Expected response rate $(\sum_h \varphi_{hk} W_h)$	Relative cost per interview	Expected number of interviews costing the same as n_0	Estimated Bias
1	0,5	1,00	1500	+1,0
2	0,849	1,12	1339	+0,7
3	0,925	1,27	1181	+0,4
4	0,955	1,51	993	+0,3
5	0,971	2,50	600	+0,2

The calculation of the expected number of interviews costing the same amount for each call-back policy (column 3) can be explained as follows: If the budget is large enough to make n_0 = 3 000 initial calls (if only one call is made per household), the expected number of interviews actually obtained in the first call is:

$$E(n_1) = 3000 \times 0{,}50 = 1500$$

For the *same total budget* and two calls being made, the expected number of interviews must be reduced to:

$$E(n_2) = \frac{1500}{1{,}12} = 1339$$

to maintain the same cost. Similarly:

$$E(n_3) = \frac{1500}{1{,}27} = 1181,$$

$$E(n_4) = \frac{1500}{1{,}51} = 993 \text{ and}$$

$$E(n_5) = \frac{1500}{2{,}50} = 600.$$

Since the relative costs at the second and later calls increase from the first call, the expected number of interviews decreases as the number of calls required by a call-back policy increases.

The *MSE*'s for the policies with 1, 2, 3, 4 or 5 calls are obtained by substituting the expected sample sizes in (4.10). For simplicity, the within-class variances are all taken as 2500. Table 4.5 represents the resulting *MSE*'s for the amount of expenditure corresponding to $n_0 = 3000$ for a single call. For comparison, the *MSE*'s are also shown for budgets corresponding to $n_0 = 1500$ and $n_0 = 2000$ for a single call.

Table 4.5 ***Values of MSE($\bar{y}$) for different call-back policies costing the same amount***

Number of calls required	n_0 =1500	n_0=2000	n_0=3000
1	4,3	3,5	2,7
2	<u>4,2</u>	<u>3,3</u>	2,4
3	4,4	<u>3,3</u>	<u>2,3</u>
4	5,1	3,9	2,6
5	8,4	6,3	4,2

The policies giving the lowest *MSE*'s in Table 4.5 are underlined. Consider first the smallest initial sample size $n_0 = 1500$. The policies requiring up to three calls produce about the same accuracy, although two calls is the optimum. For an initial sample size of $n_0 = 2000$, two to three calls are satisfactory, while four calls result in an increase in the *MSE* of approximately $\frac{3,9-3,3}{3,3} = 18,2\%$. For $n_0 = 3000$, the optimum number of calls is three. A single call results in an approximate increase in the *MSE* of $\frac{2,7-2,3}{2,3} = 17,4\%$.

This example illustrates the usefulness of accumulating information about costs and relative biases, so that an economic policy can be worked out in advance for any specific type of survey. The procedures for making call-backs and the number of calls that should be made on sample elements before classifying them as non-respondents should be made part of the survey design. The call-back policy may be modified if the results from the initial calls indicate this to be desirable. Also, the number of calls need not be the same over the entire sample but can be varied for different parts of the sample. For example, if call-backs to remote areas are much more expensive, their number may be reduced according to optimum allocation formulas.

It is quite obvious that no single optimum call-back policy can be recommended for every practical situation. Rao (1983a:43) makes the general statement that if the cost increase of later calls is moderate and the budget is sufficient to sustain an initial sample size of < 1000, about 3 or 4 calls are enough to reach the optimum but if the sample size can be increased to 2000, the optimum is reached at the fifth or later call for the same type of cost increase.

4.5. SUB-SAMPLING OF NON-RESPONDENTS

Instead of following-up all non-respondents, an alternative due to Hansen and Hurwitz (1946) is to restrict the follow-ups to a random *sub-sample* of the initial non-respondents. The restriction to a smaller sample of non-respondents allows more intense (and costly) methods of data collection to be used in the follow-ups. Hansen and Hurwitz (1946) provide the theoretical framework to determine the optimum sub-sampling fraction and the initial sample size which make the expected cost of the survey a minimum for a desired precision of the estimator. The technique was originally intended for mail surveys where the (generally) large number of non-respondents to the initial mail phase are followed-up by relatively expensive face-to-face interviews. The technique, however, can also be applied in interview surveys.

4.5.1. The Hansen and Hurwitz Procedure

4.5.1.1. Assumptions

In their original article, Hansen and Hurwitz (1946) assume that the population consists of $H=2$ mutually exclusive and exhaustive sub-groups: sub-group 1 consisting of N_r population elements who would respond if they were selected for the sample and sub-group 2 consisting of N_{nr} elements who would not respond (Rao 1983b:97). The model can also be extended to include $H \geq 2$ mutually exclusive and exhaustive sub-groups where one sub-group contains all the hard-core non-respondents and the other $H-1$ sub-groups are distinguished by the intensity of efforts required to obtain a response (Lessler & Kalsbeek 1992:178).

Using the notation and definitions in section 2.3.2.1, suppose that in the follow-up, a sub-sample of size $m = \frac{n_{nr}}{k}$ is selected from the n_{nr} non-respondents, where $\frac{1}{k}$ is the pre-determined sub-sampling fraction ($k \geq 1$). Through intensive efforts, for example face-to-face interviews, responses are obtained from all m elements in the sub-sample. (If $k = 1$, the implications are that all non-respondents are followed-up, i.e., $m = n_{nr}$, and a 100% response rate is obtained in the follow-up, so that the overall response rate is 100%.) The final sample size after the follow-up is $n_r + m$. The mean of the m respondents in the sub-sample is denoted as $\bar{y}_m = \frac{1}{m}\sum_{i=1}^{m} y_i$ and their variance as $s_m^{\ 2}$.

The existing theory presumes a 100% response rate in the sub-sample - it is this requirement that makes unbiased estimation possible. However, complete response in the sub-sample is seldom attained in practice since despite extraordinary efforts, some hard-core non-respondents will remain (Särndal *et al.* 1992:567). Some adjustments, for example weighting adjustments, may therefore have to be made to account for the hard-core non-respondents (see Chapter 5).

4.5.1.2.The Estimator and its Variance

The Hansen and Hurwitz estimator of the population mean is:

$$\hat{\bar{Y}}_{HH} = r\bar{y}_r + \tilde{r}\bar{y}_m \tag{4.11}$$

This estimator will be unbiased if responses are obtained from all the elements in the sub-sample. This means that, since the sub-sample is a *random* sample of the n_{nr} non-respondents, $\bar{y}_m$ is an unbiased estimator of $\bar{y}_{nr}$, i.e.,

$$E(\bar{y}_m \mid n_r) = \bar{y}_{nr} \tag{4.12}$$

Thus:

$$E(\hat{\bar{Y}}_{HH} \mid n_r) = r\bar{y}_r + \tilde{r}\bar{y}_{nr} = \bar{y} \tag{4.13}$$

so that:

$$E(\hat{\bar{Y}}_{HH}) = E(\bar{y}) = \bar{Y} \tag{4.14}$$

The unconditional variance of the estimator (4.11) can be written as (Rao 1983b:98):

$$\begin{aligned} V(\hat{\bar{Y}}_{HH}) &= V\left[E(\hat{\bar{Y}}_{HH} \mid n_r)\right] + E\left[V(\hat{\bar{Y}}_{HH} \mid n_r)\right] \\ &= \frac{(1-f)}{n}S^2 + \tilde{R}\frac{(k-1)}{n}S_{nr}^{\ 2} \end{aligned} \tag{4.15}$$

The first term on the right-hand side of (4.15) is the variance of the mean of a simple random sample in the case of a 100% response rate, i.e., $V(\bar{y}_{srs})$. The second term on the right-hand side of (4.15) is the increase in variance due to sub-sampling. As can be expected, this increase will be small if the proportion of non-respondents $\tilde{R}$ and their variance S_{nr}^{2} are small and the sub-sampling fraction $\frac{1}{k}$ is large (i.e., k approaches 1) (Rao 1983b:99). In other words, the Hansen-Hurwitz estimator may remove the bias from non-response, but it is less efficient than the *srs*-estimator in the case of 100% response, i.e., $V(\hat{\bar{Y}}_{HH}) \geq V(\bar{y}_{srs})$. The equality holds if $k = 1$, i.e., if all the non-respondents are followed-up and all respond. In that case, the second term on the right-hand-side of (4.15) disappears.

The variance (4.15) may be estimated by (Rao 1983b:99):

$$\begin{aligned} v(\hat{\bar{Y}}_{HH}) = {} & \frac{(N-n)(n_r-1)}{N(n-1)} r \frac{s_r^{2}}{n_r} \\ & + \frac{(N-1)(n_{nr}-1)-(n-1)(m-1)}{N(n-1)} \tilde{r} \frac{s_m^{2}}{m} \\ & + \frac{N-n}{N(n-1)}\left[r(\bar{y}_r-\hat{\bar{Y}})^2 + \tilde{r}(\bar{y}_m-\hat{\bar{Y}})^2 \right] \end{aligned} \tag{4.16}$$

Hansen and Hurwitz (1946:520) show that sending out, say, 1500 questionnaires in a survey with an expected response rate of 50% and obtaining a total of 1125 questionnaires actually processed (750 by mail and 375 by a face-to-face interview follow-up) (k = 2) yields exactly the same precision as sending out 10000 questionnaires and obtaining a total of 5263 questionnaires in the sample (5000 by mail and 263 by face-to-face interview follow-up) (k = 19). It is clear that "... at some point ... it would be unprofitable to put money into obtaining additional mail questionnaires and that it would be better to spend money on obtaining interviews from those not responding to the mail questionnaires" (Hansen & Hurwitz 1946:520).

4.5.1.3.Optimum Values of n and k

Hansen and Hurwitz (1946) consider the simple cost model:

$$c_0 n + c_1 n_r + c_2 m \tag{4.17}$$

where c_0 is the initial cost for "setting up" the survey, c_1 is the cost per element for obtaining and processing the responses from the n_r respondents and c_2 is the cost per element for contacting the sub-sampled elements and for obtaining and processing their responses. If the mail survey is followed-up by face-to-face interviews, the value of c_2 will usually be considerably larger than the value of c_1. From (4.17) the expected cost is:

$$C = \left(c_0 + c_1 R + \frac{c_2 \widetilde{R}}{k} \right) n \tag{4.18}$$

Choosing the quantities n_{opt} and k_{opt} so as to minimise the variance $V(\hat{\bar{Y}}_{HH}) \equiv V$ for a specified expected cost C or to minimise the expected cost C for a specified variance V are both equivalent to minimising the product $C\left(V + \frac{S^2}{N}\right)$ (Cochran 1977:97).

The optimum value for k is (Cochran 1977:372):

$$k_{opt} = \sqrt{\frac{c_2 (S^2 - \widetilde{R} S_{nr}^{\ 2})}{S_{nr}^{\ 2} (c_0 + c_1 R)}} \tag{4.19}$$

From (4.19) can be seen that the optimum size of the sub-sampling fraction $\frac{1}{k}$ is large (k_{opt} is small) relative to n_{nr} if c_2 is small relative to $(c_0 + c_1 R)$ or $S_{nr}^{\ 2}$ is large relative to S^2 (Rao 1983b:99). This means that if call-backs are not *much* more expensive than the original sample or if the non-respondents are more variable with respect to the survey variable than the population in general, a large sub-sampling fraction (k small) should be used, i.e., more non-respondents should be followed-up.

The initial sample size n may be chosen either to minimise the expected cost C for a specified variance $V(\hat{\bar{Y}}_{HH}) = V$ or to minimise the variance V for a specified expected cost C (Cochran 1977:372). The optimum sample size for a fixed variance V is:

$$
\begin{aligned}
n_{opt} &= \frac{N[S^2 + (k_{opt} - 1)\widetilde{R}S_{nr}{}^2]}{NV + S^2} \\
&= n_0\left[1 + \frac{(k_{opt} - 1)\widetilde{R}S_{nr}{}^2}{S^2}\right]
\end{aligned}
\tag{4.20}
$$

where $n_0 = \dfrac{NS^2}{NV + S^2}$ is the sample size required to achieve the desired variance if there is no non-response (Srinath 1971:584). The optimum sample size for a fixed expected cost C is:

$$
n_{opt} = \frac{k_{opt}C}{k_{opt}(c_0 + c_1R) + c_2\widetilde{R}}
\tag{4.21}
$$

As can be expected, the initial sample size and the sub-sampling fraction $\frac{1}{k}$ required to achieve the desired variance, will vary with the non-response rate. The solutions in (4.19), (4.20) and (4.21) therefore require a knowledge, in advance of the survey, of the non-response rate in the population ($\widetilde{R}$). This can often be estimated from previous experience or from a pilot study, but in cases where nothing is known about the non-response rate, the model can still be used as described in section 4.5.1.4.

In addition to S^2, whose value must be estimated in advance in any "optimum allocation problem", the solutions also involve $S_{nr}{}^2$, the variance in the non-response sub-group. The value of $S_{nr}{}^2$ may be harder to predict than the value of S^2 and, contrary to the assumptions made in the examples in sections 4.5.1.6 and 4.5.2.2, this variance will probably not be the same as S^2 (Cochran 1977:372). Cochran (1977:372) names one example in (mail) business surveys where the respondents tend to be larger businesses with larger between-unit-variances than the non-respondents.

4.5.1.4. Determination of the Optimum when $\widetilde{R}$ is Unknown

If there is no prior knowledge about $\widetilde{R}$, Hansen and Hurwitz (1946:522) suggest working out the value of n_{opt} from a provisional (4.19) and (4.20) for a range of assumed values of $\widetilde{R}$ between 0 and a "safe upper limit". The maximum $n_{opt} = n'$ in this series, i.e., the maximum sample size, no matter what the response rate, should then be used as the initial sample size n. When the responses to the mail survey have been received, the actual value of $n_{nr} = n' - n_r$ and hence the sample non-response rate will be known. To then determine the optimum value of k, the variance, conditional on the known values of n_{nr} and n', is used. This conditional variance can be written as (Cochran 1977:372):

$$V(\hat{\bar{Y}}_{HH} \mid n' ; n_{nr}) = \frac{(1-f)}{n'} S^2 + \frac{\tilde{r}(k-1)}{n'} S_{nr}^{\ 2} \qquad (4.22)$$

With this method, the size of the sub-sample will vary with the response rate *actually obtained* and not the *expected* response rate. Hansen and Hurwitz (1946:523) show that the cost of this method is usually only slightly higher than the optimum cost for known $\tilde{R}$ (see section 4.5.1.5).

4.5.1.5.Gains from Sub-sampling

To determine the optimum sample size and sub-sampling fraction by means of the Hansen and Hurwitz procedure, the likely non-response rate and the variability (with regard to the survey variable) in the entire population and in the non-respondent sub-group must be estimated. According to Moser and Kalton (1971:178), if reasonable estimates of these quantities can be made, the method is useful for mail surveys but the value of the method in interview surveys is more questionable.

Kish (1965:556) gives two reasons why the Hansen and Hurwitz procedure is impractical in most interview surveys: (1) the element costs of early and late calls seldom differ enough to justify introducing the "complexities of sub-sampling with its bookkeeping and weights" (see section 4.4.1), and (2) the introduction of sub-sampling into the field procedures tends to be rather awkward in many survey situations.

Durbin (1954:77) has pointed out that sub-sampling is unlikely to show a marked profit unless c_2 (the expected element cost of obtaining and processing the responses from the follow-ups) is large in relation to $c_0 + c_1 R$ (the expected element cost of obtaining and processing the responses in the first attempt). Cochran (1977:373) gives the cost ratio of obtaining a specified V in the case of k = 1 (100% follow-up of non-respondents) to the minimum cost for optimum k as:

$$\text{cost ratio} = \frac{c_0 + c_1 R + c_2 \widetilde{R}}{\left[\sqrt{R(c_0 + c_1 R)} + \sqrt{c_2}\,\widetilde{R}\right]^2} \tag{4.23}$$

if S^2 and $S_{nr}{}^2$ are approximately equal. If u denotes the ratio of the element cost of follow-ups to the initial element costs, i.e., the ratio of c_2 to $(c_0 + c_1 R)$, (4.23) can be written as:

$$\text{cost ratio} = \frac{1 + u\widetilde{R}}{\left(\sqrt{R} + \sqrt{u}\widetilde{R}\right)^2} \tag{4.24}$$

Suppose, for example, the non-response rate is 60% and the ratio of the average cost of a follow-up response to that of a first-call response is 2:1 (i.e., $u = 2$), then the ratio of the cost of obtaining a specified V in the case of $k = 1$ (100% follow-up of non-respondents) to the minimum cost for *optimum* k is:

$$\frac{1 + 2(0{,}60)}{\left(\sqrt{0{,}40} + \sqrt{2}(0{,}60)\right)^2} = 100{,}3\%$$

Thus, the relative gain in economy of sub-sampling non-respondents is only 0,3%.

Table 4.6 gives the relative gain in economy (from (4.24)) of using the optimum sub-sampling fraction $\dfrac{1}{k_{opt}}$ compared to a 100% follow-up ($k = 1$) for non-response rates of 60% and 50% and for element costs differing by factors of 2, 5, 10 and 20.

Table 4.6 ***Relative gain in economy of sub-sampling over 100% follow-ups***

$\widetilde{R}$	0,60	0,60	0,60	0,60	0,50	0,50	0,50	0,50
u	2	5	10	20	2	5	10	20
Gain in Economy	0,3%	2,6%	9,4%	18,2%	0%	5,1%	14,6%	27,0%

According to Kish (1965:536), element costs should differ by a factor of at least 4 for small savings and by perhaps 25 to obtain large savings. The figures in Table 4.6 seem to support this statement, although if S^2 is substantially greater than $S_{nr}{}^2$, there may be more to gain from sub-sampling than what is reflected in Table 4.6 (Cochran

1977:373). In face-to-face interview surveys, element costs seldom differ to a large extent, especially if the basic costs of interviewing, coding and processing are the same for both the original calls and the subsequent call-backs. On the other hand, in mail surveys with face-to-face interview follow-ups, the interviews usually cost much more than the mailed questionnaires.

4.5.1.6.Example of the Application of the Model

Suppose in a mail survey using a simple random sample design, the aim is to estimate the population mean of the survey variable y. The non-respondents to the mailed questionnaires are to be sub-sampled and interviewed in person. The optimum number of questionnaires that should be sent out and the proportion of the non-respondents who should be interviewed must be determined. The precision required is that which would be given by a simple random sample of size $n_0 = 1000$ if there was no non-response, i.e., the desired variance is $V = \frac{(N-1000)}{N}\frac{S^2}{1000}$. No advance knowledge is available about the non-response rate in the population. The cost of mailing a questionnaire is $c_0 = 75$ cents and the cost of processing a completed questionnaire is $c_1 =$ R3,00. To carry out a face-to-face interview costs $c_2 =$ R120. Suppose the variances S^2 and $S_{nr}{}^2$ are assumed equal and N is assumed to be large.

The optimum n and k (as well as the size of the sub-sample m) as calculated from (4.19) and (4.20) are given in Table 4.7 (columns 2 and 3) for expected non-response rates ranging from 10% to 90%. Also given is the cost of the optimum solution if the non-response rate were known (column 6). In column 7 the cost of an initial sample of size $n = 1000$ with a 100% interview follow-up of non-respondents ($k = 1$) is given for each non-response rate. Note that all the solutions in Table 4.7, including the 100% follow-up, give the same level of precision.

Table 4.7 Optimum solutions and expected costs that lead to the same precision for various response rates

(1) $\widetilde{R}$	(2) k_{opt}	(3) n_{opt}	(4) n_r	(5) m	(6) *Cost of optimum when $\widetilde{R}$ is known*	(7) *Cost of $n_0 = 1000$ and $k = 1$*
0,90	3,381	3143	314	837	R103 739	R109 050
0,80	4,216	3573	715	678	R86 185	R97 350
0,70	4,671	3570	1071	535	R70 091	R85 650
0,60	4,961	3377	1351	408	R55 546	R73 200
0,50	5,164	3082	1541	298	R42 695	R62 250
0,40	5,314	2726	1636	205	R31 553	R50 550
0,30	5,429	2329	1630	129	R22 117	R38 850
0,20	5,521	1904	1523	69	R14 277	R27 150
0,10	5,595	1460	1314	26	R8 595	R15 450

To illustrate how the values in Table 4.7 have been calculated, consider $\widetilde{R} = 0{,}50$. From (4.19):

$$\begin{aligned} k_{opt} &= \sqrt{\frac{c_2(1-\widetilde{R})}{c_0 + c_1 R}} \\ &= \sqrt{\frac{(120)(0{,}5)}{0{,}75 + (3{,}00)(0{,}5)}} = 5{,}164 \end{aligned}$$

assuming that $S^2 = S_{nr}^{\ 2}$ and from (4.20):

$$\begin{aligned} n_{opt} &= \frac{S^2[1 + (k_{opt} - 1)\widetilde{R}]}{V} \\ &= 1000[1 + (5{,}164 - 1)(0{,}5)] \\ &= 3082 \end{aligned}$$

where $V = \dfrac{S^2}{1000}$ is the specified desired variance, ignoring the finite population correction. Thus, if the non-response rate is 50%, the optimum solution is to send out 3082 mail questionnaires. Of the expected 1541 that are not returned, a random sub-sample of $m = \frac{n_{opt}\tilde{R}}{k_{opt}} = \frac{1541}{5,164} = 298$ should be interviewed. The expected cost of the optimum solution is from (4.18):

$$C = 0,75(3082) + 3,00(1541) + 120(298) = R42\,695.$$

The optimum values of k and n for non-response rates ranging from 0% to 100% are presented graphically in Figures 4.1 and 4.2 respectively. Figure 4.3 represents the values of k_{opt} for fixed values $c_0 = 0,75$ and $c_1 = 3,00$ but three different values of c_2, namely $c_2 = 60$, $c_2 = 120$ and $c_3 = 240$.

Figure 4.1 Optimum sub-sampling fraction for various levels of non-response

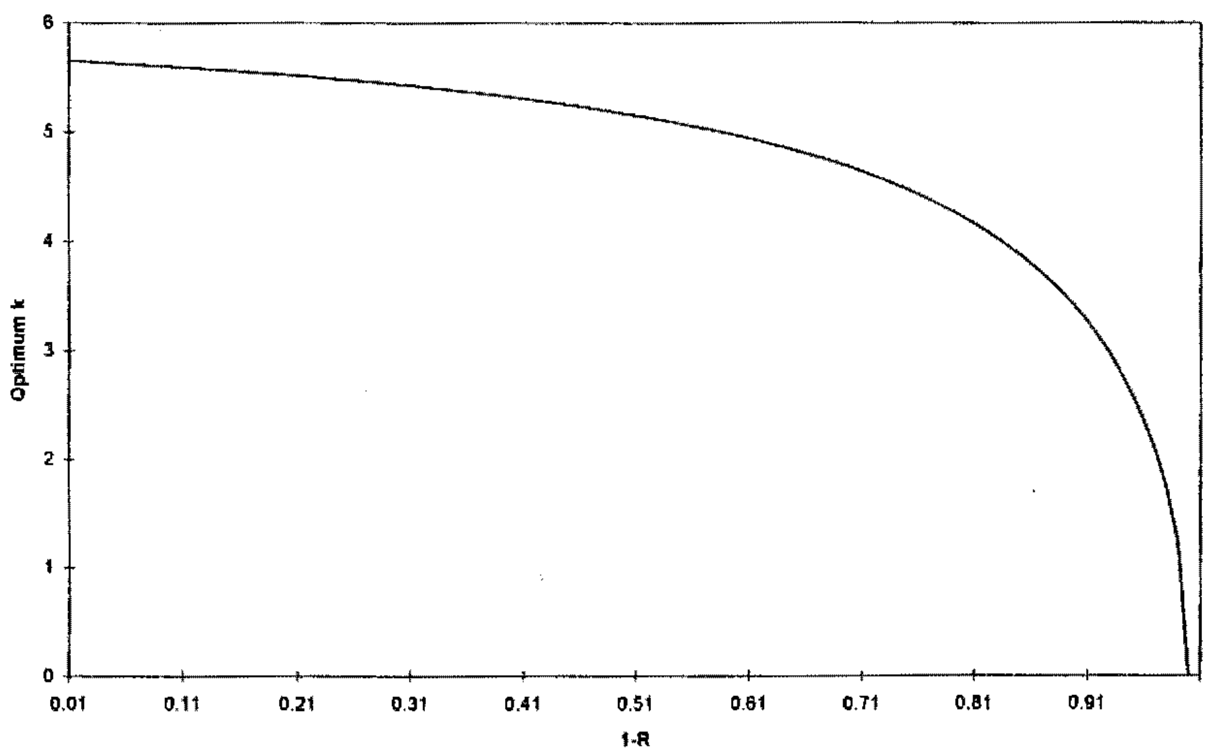


Figure 4.2 ***Optimum sample size for various levels of non-response***

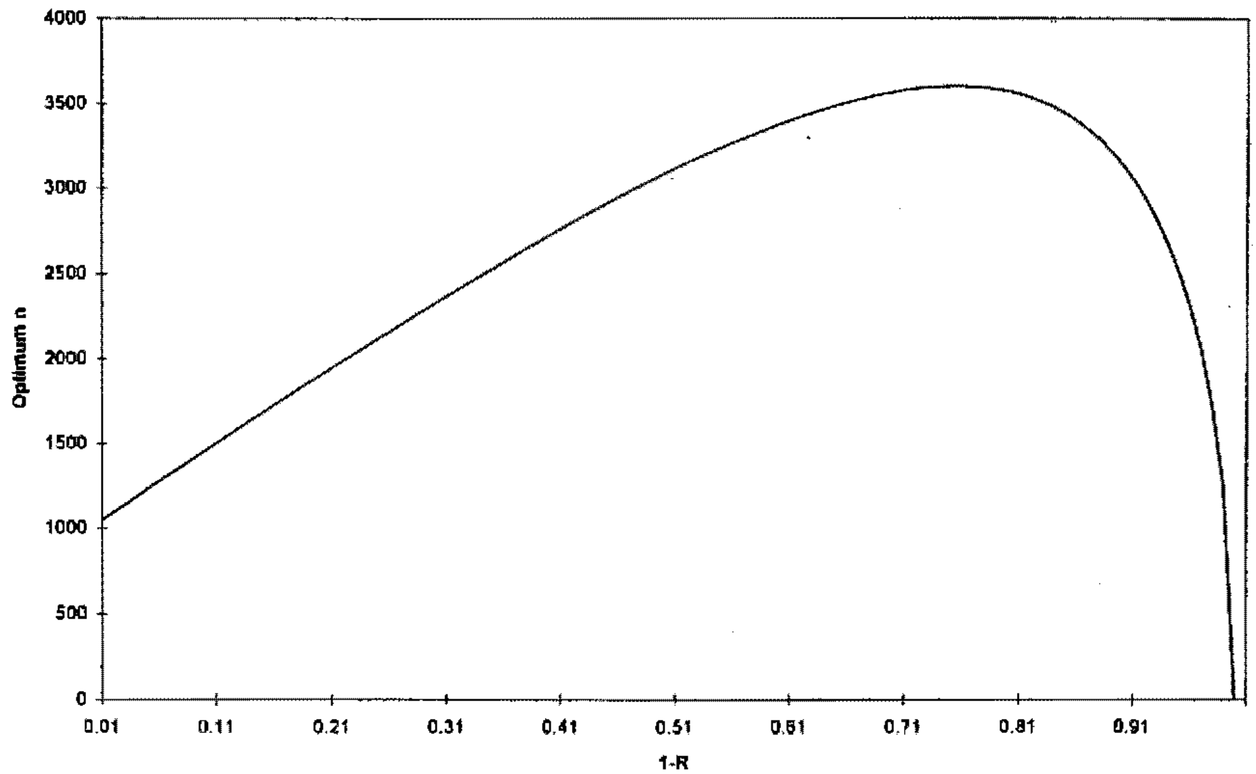
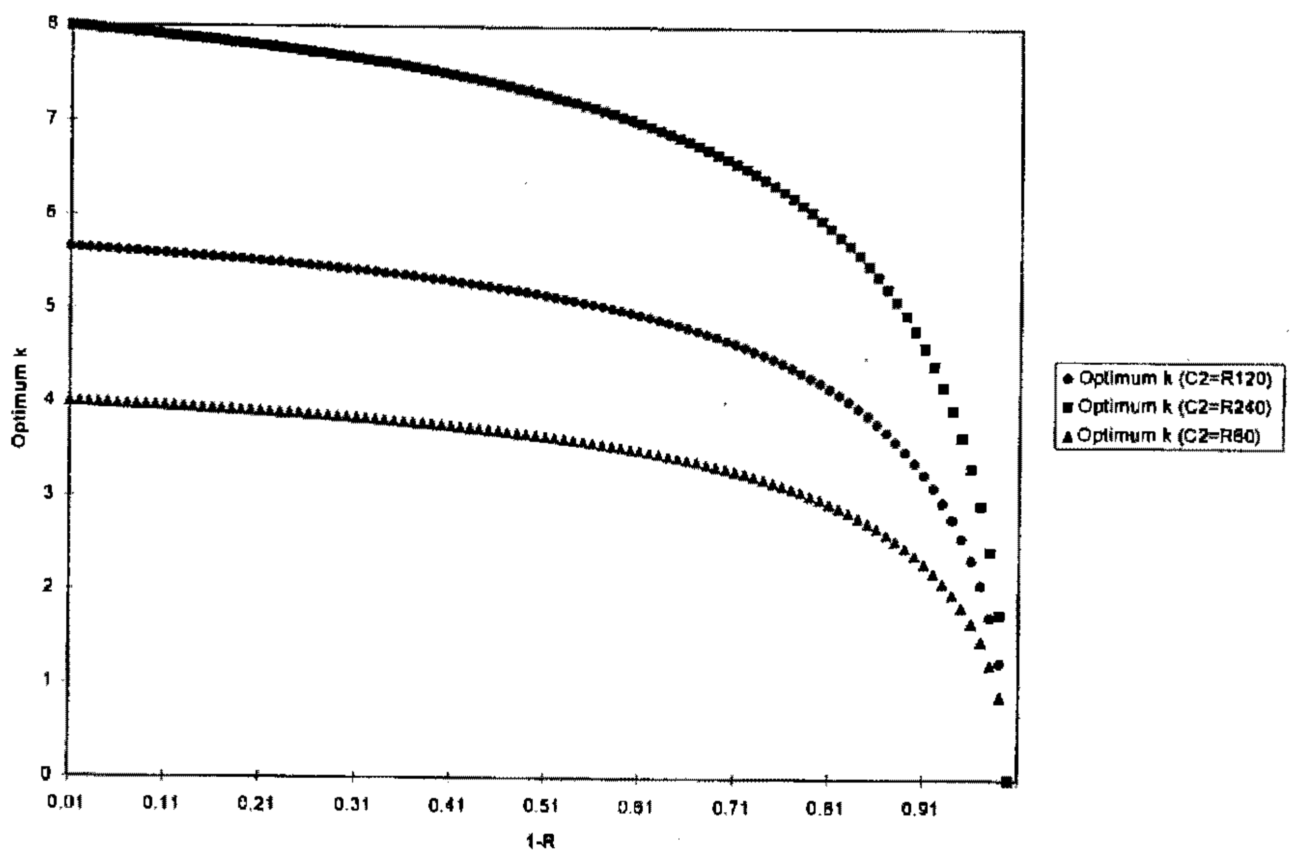


Figure 4.3 ***Optimum sub-sampling fraction for three different cost ratios***



A possible alternative to interviewing only a sub-sample of the non-respondents, is to send out 1000 questionnaires and interview all non-respondents (see paragraph 1 of section 4.5.1.6), i.e., let $k = 1$. The precision would then be $V = \frac{S^2}{1000}$, regardless of the response rate. Although this alternative gives the same precision as the optimum solution, it will always cost more than the optimum, as can be seen from column 7 of Table 4.7. For high response rates, the cost of following-up all non-respondents can be up to 90% more than the cost of following-up a sub-sample. For the very low response rates, however, a 100% follow-up does not cost much more than the optimum - which is to be expected, since not enough questionnaires have been received to take full advantage of the economies of the mail questionnaire (Hansen & Hurwitz 1946:523).

In the case of a very low response rate in the mail phase, it may be advantageous to use further mail follow-ups to maximise mail returns before sub-sampling the remaining non-respondents for interviewing. According to Claussen and Ford (1947:504):

> *... unless one has at his disposal a large, well-distributed field-staff, the time required for interviewing a larger number of non-respondents after a single mailing may be considerably greater than the time required for a mail follow-up and subsequent interviewing of fewer non-respondents.*

El-Badry (1956) has extended the Hansen and Hurwitz procedure to include several mail attempts before the face-to-face interview phase.

Note from column 3 of Table 4.7 that the maximum value of n_{opt} in this example is $n' = 3573$, no matter what the response rate. In the absence of any prior knowledge about the response rate, the optimum solution is therefore to send out 3573 questionnaires. Suppose now that the mail phase of data collection has been completed and 40% of the mail questionnaires were returned, i.e., $\tilde{R} = 0{,}60$. Using this knowledge, the optimum size of the sub-sample is determined as $m = \frac{3573(0{,}60)}{4{,}961} = 432$. The expected cost of this solution is (column 4):

$$C = 0{,}75(3573) + 3{,}00(1429) + 120(432) = R58\,807.$$

Table 4.8 shows for various *actual* response rates, the optimum number of interviews as well as the total cost if 3573 questionnaires were mailed initially. The cost of each optimum that could have been used if the response rates were known in advance is also shown as well as the costs of the 100% follow-up. (Columns 5 and 6 of Table 4.8 are identical to columns 6 and 7 of Table 4.7).

Table 4.8 Expected cost of optimum if prior knowledge of $\tilde{R}$ is lacking

(1) $\tilde{R}$	(2) k_{opt}	(3) m	(4) *Cost of optimum* ($\tilde{R}$ *unknown*)	(5) *Cost of optimum* ($\tilde{R}$ *known*)	(6) *Cost of* $n = 1000$ *and* $k = 1$
0,90	3,381	951	R117 871	R103 739	R109 050
0,80	4,216	678	R86 185	R86 185	R97 350
0,70	4,671	535	R70 096	R70 091	R85 650
0,60	4,961	432	R58 807	R55 546	R73 200
0,50	5,164	346	R49 561	R42 695	R62 250
0,40	5,314	269	R41 392	R31 553	R50 550
0,30	5,429	197	R33 823	R22 117	R38 850
0,20	5,521	130	R26 854	R14 277	R27 150
0,10	5,595	64	R20 008	R8 595	R15 450

From a comparison of columns 4 and 5 it can be seen that, if the response rate is known approximately in advance, the use of this information in determining the optimum solution, may lead to lower cost depending on the value of $\tilde{R}$. However, for the very high response rates, the lack of any advance knowledge of the response rate entails almost no additional cost over the optimum value when the rate is known in advance. This can be expected, since, when the response rates are high, the total cost of the survey will be small even though an unnecessarily large number of questionnaires had originally been sent out (Hansen & Hurwitz 1946:525).

From this example it can be seen that the Hansen and Hurwitz procedure can be used to find optimum values of n and k not only when the response rate is known in advance, but also when nothing is known about the rate of response and this procedure will produce results having at least the specified precision and at lower cost than a 100% follow-up of non-respondents.

4.5.2. An Alternative Procedure by Srinath (1971)

Srinath (1971) suggested an alternative to the Hansen and Hurwitz procedure for obtaining the value of the optimum sub-sampling fraction among the non-respondents. This method does not require any advance knowledge of $\widetilde{R}$. The sub-sampling fraction is not fixed as in the Hansen and Hurwitz procedure, but is allowed to vary according to the observed non-response rate. The variance of the estimator of the population mean is consequently independent of the unknown non-response rate in the population (Srinath 1971:583).

4.5.2.1. Optimum Values of n and k

Srinath (1971) suggests determining the size of the sub-sample as:

$$m' = \frac{n_{nr}^{2}}{k'n + n_{nr}} = n\frac{\widetilde{r}^{2}}{k' + \widetilde{r}} = m\frac{k\widetilde{r}}{k' + \widetilde{r}} \tag{4.25}$$

where $k' > 0$ is "some constant fixed in advance", k is the constant determined by the Hansen and Hurwitz procedure and $m = \frac{n}{k}$.

It can be shown that (Rao 1983b:100):

$$\begin{aligned} V'(\hat{\bar{Y}}_{HH} \mid n_r) &= \widetilde{r}^{2}\left(\frac{1}{m'} - \frac{1}{n_{nr}}\right)s_{nr}^{2} \\ &= \frac{k'}{n}s_{nr}^{2} \end{aligned} \tag{4.26}$$

so that the unconditional variance of the estimator can be written as:

$$\begin{aligned} V'_{HH}(\hat{\bar{Y}}) &= V[E(\hat{\bar{Y}}_{HH} \mid n_r)] + E[V'(\hat{\bar{Y}}_{HH} \mid n_r)] \\ &= \frac{(1-f)}{n}S^{2} + \frac{k'}{n}S_{nr}^{2} \end{aligned} \tag{4.27}$$

For this method of choosing the size of the sub-sample, the cost function analogous to (4.17) is:

$$c_0 n + c_1 n_r + c_2 m' \tag{4.28}$$

and from (4.28) the *approximate* expected cost is (Srinath 1971:584):

$$C' = (c_0 + c_1 R)n + c_2\left(\frac{n\widetilde{R}^2}{k' + \widetilde{R}}\right) \tag{4.29}$$

which is obtained by replacing $\widetilde{r}$ and $\widetilde{r}^2$ by $\widetilde{R}$ and $\widetilde{R}^2$.

From (4.25) can be seen that m is larger or smaller than m' as k' is larger or smaller than $(k-1)\widetilde{r}$. Similarly, it can be seen from (4.15) and (4.27) that $V_{HH}(\hat{\bar{Y}})$ is smaller or larger than $V'_{HH}(\hat{\bar{Y}})$ and from (4.18) and (4.29), that C is larger or smaller than C' as k' is larger or smaller than $(k-1)\widetilde{R}$ (Rao 1983b:100).

Minimising C' for a given value of V or minimising V for a given value of C', the solution to the optimum value of k' can be obtained as:

$$k'_{opt} = \sqrt{\frac{\left(S^2 - \widetilde{R}S_{nr}^{\ 2})c_2\widetilde{R}^2\right)}{S_{nr}^{\ 2}(c_0 + c_1 R)}} - \widetilde{R} \tag{4.30}$$

which can be written in terms of the Hansen and Hurwitz k_{opt} as:

$$k'_{opt} = (k_{opt} - 1)\widetilde{R} \tag{4.31}$$

For a specified value V of the variance:

$$n'_{opt} = n_0\left[1 + k'_{opt}\left(\frac{S_{nr}^{\ 2}}{S^2}\right)\right] \tag{4.32}$$

and for a specified C':

$$n'_{opt} = \frac{C'(k'_{opt} + \widetilde{R})}{(c_0 + c_1 R)(k'_{opt} + \widetilde{R}) + c_2\widetilde{R}^2} \tag{4.33}$$

Equations (4.30) and (4.32) lead to the same initial sample size n_{opt} and expected cost C as the Hansen-Hurwitz procedure.

4.5.2.2. Example of the Srinath Method

To compare the Hansen and Hurwitz procedure to the Srinath procedure, consider the example in section 4.5.1.5 in which $c_0 = 0{,}75$; $c_1 = 3$ and $c_2 = 120$. It is again assumed that S^2 and $S_{nr}^{\ 2}$ are equal. If the value of $\widetilde{R}$ is thought to be equal to 0,4, $k_{opt} = 5{,}314$ from Table 4.7. From (4.31):

$$
\begin{aligned}
k'_{opt} &= (k_{opt} - 1)W_{nr} \\
&= (5{,}314 - 1)0{,}40 \\
&= 1{,}726
\end{aligned}
$$

and from (4.20) and (4.32):

$$
\begin{aligned}
n_{opt} &= n'_{opt} = n_0\left[1 + k'_{opt}\right] \\
&= 1000(1 + 1{,}726) \\
&= 2726
\end{aligned}
$$

The optimum size of the sub-sample for the Hansen and Hurwitz method is:

$$m = \frac{n_{opt}\widetilde{R}}{k_{opt}} = \frac{2726(0{,}4)}{5{,}314} = 205 \tag{4.34}$$

and for the Srinath procedure:

$$m' = n_{opt}\frac{\widetilde{R}^2}{k'_{opt} + \widetilde{R}} = \frac{2726(0{,}16)}{1{,}726 + 0{,}4} = 205 \tag{4.35}$$

For the present case, both m and m' are equal to 205. From (4.18) and (4.29) the expected cost is $C = 0{,}75(2726) + 3{,}00(1636) + 120(205) = R31553$.

But suppose that the true value of $\widetilde{R}$ is equal to 0,6 instead of 0,4 as was thought earlier. With the earlier computed values of k_{opt} of 5,314 and k'_{opt} of 1,726 and this value 0,6 for $\widetilde{R}$, the precision of the Hansen and Hurwitz procedure is, from (4.15):

$$
\begin{aligned}
V_{HH}(\hat{\bar{Y}}) &= \frac{(1-f)}{n}S^2 + \widetilde{R}\frac{(k-1)}{n}S_{nr}^{\,2} \\
&= \left(1 - \frac{2726}{N}\right)\frac{S^2}{2726} + \frac{0{,}6(4{,}314)}{2726}S^2 \\
&\approx \frac{1}{1000}(1 + 0{,}3164)S^2 \\
&= V + \frac{0{,}3164}{1000}S^2 \quad > V
\end{aligned}
$$

but the precision of the Srinath method is, from (4.27):

$$
\begin{aligned}
V'_{HH}(\hat{\bar{Y}}) &= \frac{(1-f)}{n}S^2 + \frac{k'}{n}S_{nr}^{\ 2} \\
&= \left(1-\frac{2726}{N}\right)\frac{S^2}{2726} + \frac{1{,}726}{2726}S^2 \\
&\approx \frac{S^2}{1000} = V
\end{aligned}
$$

The Srinath procedure therefore provides the desired precision. However, if the costs of these methods are compared, from (4.18):

$$
C = \left(0{,}75 + 3(0{,}4) + \frac{120(0{,}6)}{5{,}314}\right)2726 = R42251
$$

and from (4.29):

$$
C' = [0{,}75 + 3(0{,}4)]2726 + 120\left(\frac{2726(0{,}6)^2}{1{,}726 + 0{,}6}\right) = R55945\,.
$$

Thus, with an *extra* cost, the Srinath procedure would provide the required precision, although $\widetilde{R}$ is equal to 0,6 instead of 0,4.

With the increased budget of R55945 and $\widetilde{R}$ = 0,6, from Table 4.7, k_{opt} = 4,961 and from (4.21):

$$
\begin{aligned}
n_{opt} &= \frac{k_{opt}C}{k_{opt}(c_0 + c_1 R) + c_2\widetilde{R}} \\
&= \frac{4{,}961(55945)}{4{,}961[0{,}75 + 3(0{,}4)] + 120(0{,}6)} \\
&= 3398
\end{aligned}
$$

With these values $V_{HH}(\hat{\bar{Y}}) = V$. Thus, the Hansen and Hurwitz procedure also provides the required precision with the increased budget; however, it cannot be implemented since it was not known initially that $\widetilde{R}$ = 0,6 (Rao 1983b:101).

The Srinath procedure will not always provide a higher precision than the Hansen and Hurwitz method. Suppose for example that $\widetilde{R}$ = 0,1 instead of 0,4. The precision and costs of the two methods are compared in Table 4.9.

Table 4.9 ***Comparison of precision and cost of Hansen and Hurwitz procedure to Srinath procedure***

H&H	$k_{opt} = 5{,}595$	$n_{opt} = 1460$	$V_{HH}(\hat{\bar{Y}}) = V - \frac{0{,}0003}{1000}S^2 < V$	$C = \text{R}8168$
Srinath	$k'_{opt} = 0{,}460$	$n'_{opt} = 1460$	$V'_{HH}(\hat{\bar{Y}}) = V$	$C' = \text{R}8166$

Thus, with an extra cost, the Hansen and Hurwitz procedure gives more precision than required.

4.5.3. Conclusion

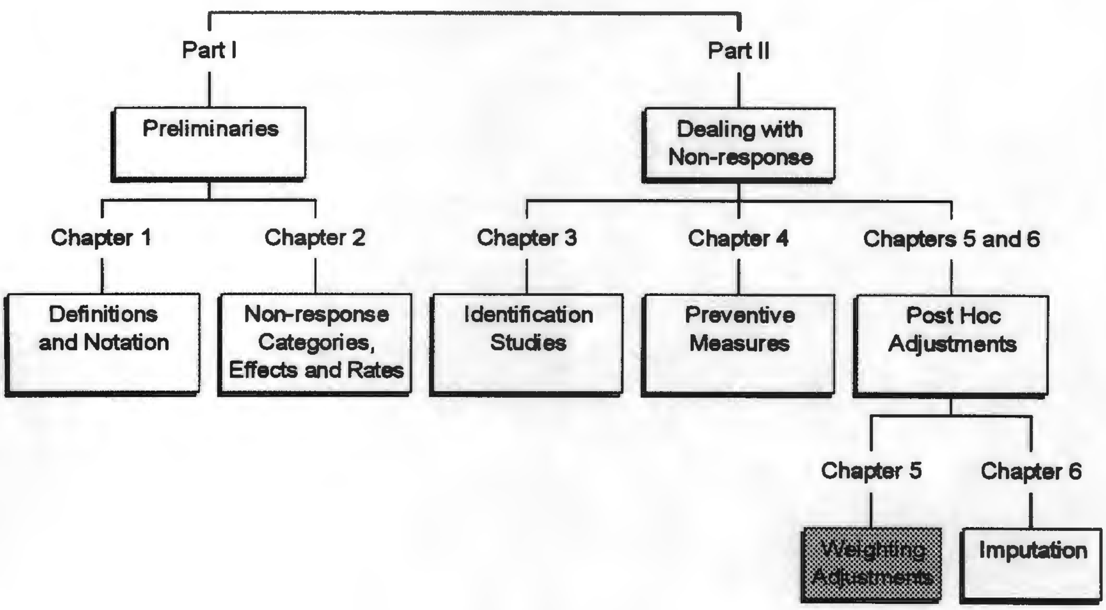
The final number of respondents in the survey will usually be lower if the technique of sub-sampling is applied than with a call-back strategy. Furthermore, as was shown, the Hansen-Hurwitz estimator is less efficient than the simple random sample estimator under a 100% response rate. However, if the sub-sampling fraction is reasonably large and if the response rate in the sub-sample is high, then sub-sampling may, nevertheless, reduce the risk of non-response biases sufficiently to compensate more than enough for the increase in variances.

In the following two chapters, the focus is on statistical methods of dealing with non-response *after* data collection is completed. However, these methods will not fully eliminate the effects of non-response. Thus, according to Madow, Nisselson and Olkin (1983a:7) the best recommendation on dealing with non-response is to collect the survey data as fully and accurately as possible, using call-backs and follow-up techniques to increase response levels as much as possible. Also, careful and appropriate preparation for data collection with respect to methods of interviewing and motivation of respondents and interviewers will considerably affect the magnitude of response.

CHAPTER 5

COMPENSATING FOR UNIT NON-RESPONSE:

WEIGHTING ADJUSTMENTS



CHAPTER OUTLINE

5.1 INTRODUCTION

5.2 THE UNIFORM GLOBAL RESPONSE MECHANISM

5.3 UNIFORM RESPONSE MECHANISM WITHIN SUB-POPULATIONS

5.4 FORMATION OF WEIGHTING CLASSES

5.5 THREE SIMPLE EXAMPLES

5.6 RAKING RATIO ESTIMATION

5.7 OTHER WEIGHTING TECHNIQUES

5.8 CONCLUSION

CHAPTER 5

COMPENSATING FOR UNIT NON-RESPONSE:

WEIGHTING PROCEDURES

... when survey sampling was in its infancy, most theorists were practitioners and most practitioners were theorists. Problems encountered in practice fostered theory, and advances in theory nurtured practice. This happy state of affairs does not seem to exist now. Today, sampling theorists do not engage in conducting surveys, and practitioners ignore the new theoretical exhortations. Can anything be done to bring theory and practice together again?

N. Krishnan Namboodiri, Survey Sampling and Measurement

5.1. INTRODUCTION

In Chapter 4, the methods of dealing with non-response were classified into two categories, namely *preventive* methods and *post hoc* methods. The various preventive methods (discussed in Chapter 4) entail attempts to collect the data as fully as possible; post hoc methods (Chapters 5 and 6) make do with the data but apply an adjustment technique to compensate for the missing data. Post hoc adjustment procedures can be separated into two types, namely *weighting adjustments* and *imputation*. Weighting adjustments increase the weights (ω) of specified respondents to compensate for the non-respondents while imputation techniques insert values for missing responses. Imputation techniques will be discussed in Chapter 6.

5.1.1. Choice Between Weighting and Imputation

The dominant factor determining the choice between these two methods for handling a particular type of missing data, e.g., unit or item non-response[1], is the amount of information available on the sample elements involved (Kalton & Kasprzyk 1986:1). In general, the only information available about *unit* non-respondents is that contained in the sampling frame, for example, the strata and/or the PSU's in which they are located. This information can usually readily be incorporated into weighting adjustments so that as a rule, *unit* non-response is compensated for by some form of weighting adjustment.

[1] Weighting may also be used to deal with missing data due to *under-coverage* (not discussed in this dissertation).

In the case of item non-response, a great deal of additional information is usually available for the sample elements involved: not only the information contained in the sampling frame but also their responses to other survey items. Item non-response is therefore usually handled by some form of imputation which attempts to incorporate all the actual responses to survey items into the compensation procedure. However, the choice between these two methods is not always clear-cut: in some surveys, the sampling frame may contain a large amount of information on unit non-respondents in which case imputation may be more appropriate than weighting. On the other hand, a substantial amount of data may be missing for a respondent who terminates an interview at an early stage or who refuses to answer most of the questions, so that weighting may be more appropriate than imputation. (Kalton 1983a:6.)

Kalton (1983a:19) shows that the techniques of weighting adjustment and imputation are actually closely related: there exists an equivalent imputation procedure for any weighting adjustment with integer weights. For example, when estimating population means and totals, weighting adjustments implicitly impute mean values for each non-respondent, similar to the technique of mean value imputation (see sections 6.2.3 and 6.2.4.) This imputation process is implicit rather than explicit since imputed values never physically replace the missing responses in the data set (Lessler & Kalsbeek 1992:212). Whereas with weighting adjustments, the records for the respondent and the non-respondent are merged into a single record with increased weight, in imputation the two records, although identical, remain separate (see Chapter 6) (Kalton 1983a:19).

5.1.2. Response Mechanisms

As stated in Chapter 2, whenever there is non-response in a survey, explicit model assumptions must be made about the characteristics of the non-respondents in order to produce (presumably) unbiased estimates of the population values of interest. These assumptions usually involve specifying a particular *response mechanism* in addition to the probability sampling mechanism.

Two alternative response mechanisms that can be considered in non-response adjustment procedures are the *uniform global response mechanism* and the *uniform response mechanism within sub-populations*, both of which can be specified as *probability* response mechanisms (see section 2.3.1). Under the uniform global (probability) response mechanism, all elements in the population, if they were selected for the survey, are presumed to have equal, positive and independent response probabilities. Under the uniform (probability) response mechanism within sub-populations, all elements in specified sub-populations are presumed to have equal, positive and independent response probabilities but the response probabilities may vary among the sub-populations.

Two alternative assumptions are possible under either of the uniform response mechanisms: (1) the data are missing at random across the entire sample (or across the specified sub-populations) which implies that the respondents are a random sample from the population (or from the sub-populations), or (2) the respondent and non-respondent means in the population (or in the specified sub-populations) are equal. As stated in Chapter 2, assumption (1) is the preferred assumption under the quasi-randomisation approach to inference. Kalton (1983a:15) shows that these two assumptions are not equivalent:

1. If the data are assumed to be missing at random, it does not necessarily imply that $\overline{Y}_r = \overline{Y}_{nr}$, although their *expected* values are presumed to be equal, i.e., $E(\overline{Y}_r) = E(\overline{Y}_{nr})$. (The notation in section 2.3.2.1. will be used throughout this chapter.)
2. Under the assumption of data missing at random, the implication is that the expected distributions of the respondents and non-respondents and hence also the expected element variances and other parameters of the distributions are equal, e.g., $E(S_r^2) = E(S_{nr}^2) = S^2$. On the other hand, if the assumption is that $\overline{Y}_r = \overline{Y}_{nr}$, these expected parameters of the respondent and non-respondent distributions are not necessarily equal.

In section 5.2, non-response adjusted estimators of the population mean and total are discussed under the uniform global response mechanism. When the uniform global response mechanism is assumed, a *constant weight* is applied to the sample as a whole. It is shown that non-response adjustment under the uniform global response mechanism actually reduces to a "do nothing" situation. The uniform global response mechanism is therefore not really useful in practice and it is discussed here mainly for comparative purposes.

When the response mechanism is assumed to be uniform within sub-populations, different weights are applied in different sub-groups (called weighting classes) of the sample. Non-response adjusted estimators of the population mean and total in the case of the uniform response mechanism within sub-populations are discussed in section 5.3. Either *sample* weighting adjustments or *population* weighting adjustments may be applied. Sample weighting adjustments are discussed in section 5.3.3 while population weighting adjustments are discussed in section 5.3.4.

In section 5.4, a few guidelines are given for the formation of weighting classes and in section 5.5, three examples are given to illustrate the techniques of population and sample weighting adjustments. Section 5.6 considers a more sophisticated but less familiar technique of adjusting for non-response, namely *raking ratio estimation*. Various other adjustment procedures employing weights of some kind are discussed in section 5.7, including integer and duplication weighting, the Politz-Simmons technique, and linear regression estimation.

5.1.3. Conditional Vs Unconditional Inferences

Various authors, such as Thomsen (1973), Oh and Scheuren (1983), Kalton (1983a), Lehtonen and Pahkinen (1995) and Little (1986) discuss weighting adjustments for non-response and give bias and variance expressions of estimators of the population mean and/or total under these weighting adjustments. However, comparisons between expressions given by these authors are complicated by different assumptions. For example, Thomsen (1973), Kalton (1983a), Lehtonen and Pahkinen (1995) and Little (1986) give moments of the estimators over the sampling distribution, conditional on the values of y, n and n_r and conditional on the response mechanism. On the other hand, Oh and Scheuren (1983) postulate a Bernoulli response distribution and calculate moments of estimators over both the sampling and the response distribution with (a) y held fixed and (b) y, n and n_r held fixed. They call the former unconditional moments and the latter conditional moments. Furthermore, Oh and Scheuren (1983) and Lehtonen and Pahkinen (1995) include finite population corrections while Thomsen (1973) and Kalton (1983a) assume that N is large and therefore ignore the finite population corrections.

Kalton (1983a) gives conditional variance expressions in the case of (a) the uniform global response mechanism and (b) population weighting adjustments with y, n and n_r held fixed (sections 5.2 and 5.3.4). However, for *sample* weighting adjustments (section 5.3.3), Kalton (1983a:51) believes it is inappropriate to take n_r as fixed since "the value of n_r depends on the sizes of the sub-populations which appear in the estimator and which are subject to sampling variability". He therefore gives unconditional variance expressions for sample weighted estimators.

Little (1986:144) states that he prefers the expressions given by Oh and Scheuren (1983) that condition on y, n and n_r, since they provide more precise results when the respondent sample sizes are small. On the other hand, he prefers, like Thomsen (1973) and Kalton (1983a), to calculate moments conditional on the response mechanism since the validity of the Oh and Scheuren (1983) calculations is specific to a particular choice of sub-populations.

As stated in Chapter 1, the approach of Holt and Smith (1979:34) will be followed in this dissertation. They argue that unconditional variances should be used when comparing sampling strategies *before* the sample is drawn and for inference *after* the sample is drawn, conditional variances are appropriate. Furthermore, according to the design-based approach (as followed in this dissertation), all inferences in this dissertation are conditional on y. As stated in section 1.2.3, all inferences in this dissertation are also conditional on the specified sample size n. In this chapter, inferences will also be made conditional on both the response mechanism and the value of n_r. Unconditional variance expressions will be given in only a few cases in this chapter to facilitate comparisons between the various adjustment techniques.

5.2. THE UNIFORM GLOBAL RESPONSE MECHANISM: Constant Weighting Adjustments

If a uniform global (probability) response mechanism is assumed, i.e., if it is assumed that $\varphi_j = \varphi$ where $0 < \varphi \leq 1$ for all $j = 1, \ldots, N$, a constant weight ω is applied to the data from each respondent to compensate for non-response.

5.2.1. Generalised Non-response Adjusted Estimators

If the uniform global (probability) response mechanism is assumed, generalised π^*-estimators (see section 1.3.1.1) of population values can be constructed by using the (known) inclusion probabilities π_i and response probabilities φ_i to construct weights $\omega_i = \dfrac{1}{\pi_i \varphi_i}$ (Särndal *et al.* 1992:558). Under the uniform global response mechanism, the value of $\varphi_i = \varphi > 0$ may be obtained as (Lessler & Kalsbeek 1992:183):

$$\varphi_i = \frac{\sum_{i=1}^{n_r} \frac{1}{\pi_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i}} \tag{5.1}$$

for $i = 1, \ldots, n_r$.

Strictly speaking, the value of φ_i calculated from (5.1) is an *estimated* response probability $\hat{\varphi}_i$: it is an unbiased estimate of the mean response probability in the population, namely $\overline{\varphi} = \dfrac{1}{N}\sum_{j=1}^{N} \varphi_j$ (Bethlehem 1988:254). In this dissertation, a distinction will not be made between a "true" response probability and its estimate. The reasons are twofold:

1. According to Lessler and Kalsbeek (1992:138) the concept of an individual response probability for each element in the population is rather difficult to envisage and it is impossible to know precisely the chances that a population element, if selected, would be interviewed and would provide useful data. Therefore, response probabilities can, at best, only be estimated or conjectured.
2. For most methods of estimating φ the effect on the *MSE* of estimators when using the estimate rather than the actual value of φ is unknown.

Using the value of φ_i obtained from (5.1), the generalised (non-response adjusted) π^*-estimator of the population total can be written as:

$$\hat{Y}_{\pi^*} = \sum_{i=1}^{n_r} \omega_i y_{r_i} = \sum_{i=1}^{n_r} \frac{y_{r_i}}{\pi_i \varphi_i} \tag{5.2}$$

where $\omega_i = \dfrac{1}{\pi_i \varphi_i}$. The generalised (non-response adjusted) π^*-estimator of the population mean can be written as:

$$\hat{\bar{Y}}_{\pi^*} = \frac{\sum_{i=1}^{n_r} \omega_i y_{r_i}}{\sum_{i=1}^{n_r} \omega_i} = \frac{\sum_{i=1}^{n_r} \frac{y_{r_i}}{\pi_i \varphi_i}}{\sum_{i=1}^{n_r} \frac{1}{\pi_i \varphi_i}} \tag{5.3}$$

5.2.2. Non-response Adjusted Estimator of the Population Total Assuming srs wor

Assuming simple random sampling without replacement or any other epsem sample of size n from the finite population, the inclusion probabilities are $\pi_i = \dfrac{n}{N}$ for all $i = 1, \ldots, n$. From (5.1) (Särndal *et al.* 1992:579):

$$\varphi_i = \frac{n_r}{n} = r \text{ for all } i = 1, \ldots, n_r \tag{5.4}$$

The constant weighted estimator of the population total is from (5.2):

$$\hat{Y}_w = \sum_{i=1}^{n_r} \omega_i y_{r_i} \tag{5.5}$$

where $\omega_i = \dfrac{N}{n} \times \dfrac{n}{n_r} = \dfrac{N}{n_r}$. Hence, (5.5) can be written as:

$$\hat{Y}_w = \frac{N}{n_r} \sum_{i=1}^{n_r} y_{r_i} = N\bar{y}_r \tag{5.6}$$

All respondents are assigned the same weight of $\omega_i \propto \dfrac{n}{n_r}$, proportional to the inverse of the sample response rate. The expected value of (5.6) over repeated simple random samples is:

$$E(\hat{Y}_w \mid n_r) = N\bar{Y}_r \tag{5.7}$$

Under the model of data missing at random across the entire sample, $E(\bar{Y}_r) = \bar{Y}$, hence, the estimator (5.6) is approximately unbiased under the model.

The conditional variance of (5.6), provided that $n_r \geq 2$, can be written as (Oh & Scheuren 1983:148):

$$\begin{aligned} V(\hat{Y}_w \mid n_r) &= N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_r^{\,2} + N^2\left(\frac{1}{n_r}-\frac{1}{n}\right)S_r^{\,2} \\ &= N^2\left(1-\frac{n_r}{N}\right)\frac{S_r^{\,2}}{n_r} \end{aligned} \tag{5.8}$$

Under the model of data missing at random, $S_r^{\,2}$ is an unbiased estimator of S^2 (Oh & Scheuren 1983:148). If $S_r^{\,2}$ in (5.8) is replaced with its expected value S^2, the first term on the right-hand side of (5.8) is the variance of a *srs wor* estimator of the population total when there is no non-response; the second component is directly attributable to the additional level of "sampling" introduced by the response mechanism (Oh & Scheuren 1983:148).

Since $s_r^{\,2}$ is an unbiased estimator of $S_r^{\,2}$, (5.8) can be estimated by:

$$v(\hat{Y}_w \mid n_r) = N^2\left(1-\frac{n_r}{N}\right)\frac{s_r^{\,2}}{n_r} \tag{5.9}$$

5.2.3. Non-response Adjusted Estimator of the Population Mean Assuming srs wor

The application of the constant weight ω to estimate the population *mean* has no effect, since it results in the unweighted respondent mean $\bar{y}_r$. From (5.3):

$$\hat{\bar{Y}}_w = \frac{\sum_{i=1}^{n_r}\omega_i y_{ri}}{\sum_{i=1}^{n_r}\omega_i} = \frac{\frac{n}{n_r}\sum_{i=1}^{n_r} y_{ri}}{\frac{n}{n_r}n_r} = \frac{1}{n_r}\sum_{i=1}^{n_r} y_{ri} = \bar{y}_r \tag{5.10}$$

The expected value of the respondent mean over repeated samples is:

$$E(\bar{y}_r) = \bar{Y}_r \tag{5.11}$$

If the data are missing at random across the entire sample, $E(\bar{Y}_r) = \bar{Y}$ and the estimator (5.10) is approximately unbiased.

5.2.4. Discussion

The global uniform response mechanism (sections 5.2.2 and 5.2.3) has little practical utility (Oh & Scheuren 1983:148). There are a number of reasons for this:

1. The use of the constant weighted estimator of the population *mean* (5.10) is equivalent to doing nothing, i.e., simply using the respondent data to estimate the population mean.

2. When estimating the population *total*, doing nothing will result in the *biased* estimator of the population total: $\hat{Y} = \frac{N}{n}\sum_{i=1}^{n_r} y_{r_i}$ (see section 2.3.2.2.2). Nevertheless, it is *easier* to calculate the unbiased estimator of the population total as $N\bar{y}_r$ than to calculate it in the form of a weighted estimator $\hat{Y}_w = \sum_{i=1}^{n_r} \omega_i y_{r_i}$, although $\hat{Y}_w$ and $N\bar{y}_r$ will yield the same estimate. (Kalton 1983a:42.)

3. The uniform global response mechanism seldom holds in practice: as was shown in Chapter 3, the non-respondents are rarely a *random* sample from the population.

A natural extension of the uniform *global* response mechanism is to divide the population into sub-populations and then to apply the assumption of uniform response probabilities within each sub-population. Numerous surveys have provided ample evidence to demonstrate that response rates vary across sub-populations and that the survey variables are often associated with the characteristics of these sub-populations (Kalton 1983a:16).

5.3. UNIFORM RESPONSE MECHANISM WITHIN SUB-POPULATIONS

The assumption of a uniform response mechanism within sub-populations requires that the respondents or the entire sample be divided into H mutually exclusive sub-populations - called *weighting classes* or *adjustment cells*, based on suitable auxiliary variables (see section 5.4). Two types of weighting adjustments are possible, namely *sample weighting adjustments* and *population weighting adjustments*. The choice between these depends upon whether or not the population sizes of the weighting classes are known. If the distribution of the population over the weighting classes is known, say, from external sources such as previous census data, population weighting adjustments may be applied. In this case, only *respondents* need to be divided into weighting classes. If the population distribution over the weighting classes is unknown, sample weighting adjustments are applied. In this case, both respondents and non-respondents must be divided into weighting classes. The formation of weighting classes for *sample* weighting adjustments is therefore limited to auxiliary variables whose values are available for both respondents and non-respondents. This essentially restricts the characteristics by which weighting classes for sample weighting adjustments can be defined, to variables such as geographic location, race, gender, level of urbanisation, housing unit characteristics and design variables. (Kalton 1983a:50.)

A further difference between population weighting adjustments and sample weighting adjustments is that the former compensate for both under-coverage and unit non-response, while the latter compensate only for unit non-response.

Implicit in the formation of weighting classes for population and sample weighting adjustments are (Bailey 1983:291):

1. There is "significant" correlation between the principal survey variable(s) and the auxiliary variables used to define the weighting classes
2. The data are missing at random within each weighting class
3. The respondent means differ among the weighting classes (see point 2 in section 5.3.6)

Population weighting adjustments bear a close resemblance to post-stratification and are sometimes called post-stratification adjustments. However, there are important differences between weighting classes formed for non-response adjustments and post-strata formed for post-stratification (see section 5.3.4.3).

5.3.1. Notation

The notation in section 2.3.2.1 must be extended to take into account the H mutually exclusive weighting classes. Let $W_h = \frac{N_h}{N}$ be the proportion of the h-th class in the population[2]. If W_h is unknown, it is estimated by $w_h = \frac{n_h}{n}$, the proportion of the h-th class in the sample. The N_h elements in class h consist of N_{r_h} respondents and N_{nr_h} non-respondents. The proportions of respondents and non-respondents in class h are denoted respectively as $R_h = \frac{N_{r_h}}{N_h}$ and $\widetilde{R}_h = \frac{N_{nr_h}}{N_h}$. The sample size in class h is denoted as n_h and the number of sample respondents and non-respondents in class h are respectively denoted as n_{r_h} and n_{nr_h}. The sample response and non-response rates in class h are respectively denoted as $r_h = \frac{n_{r_h}}{n_h}$ and $\widetilde{r}_h = \frac{n_{nr_h}}{n_h}$. The population variance of the respondents in the h-th weighting class is defined as:

$$S_{r_h}{}^2 = \frac{1}{N_{r_h} - 1}\sum_{j=1}^{N_{r_h}}\left(y_{r_h j} - \bar{Y}_{r_h}\right)^2 \tag{5.12}$$

The sample variance of respondents in the h-th weighting class is defined as:

$$s_{r_h}{}^2 = \frac{1}{n_{r_h} - 1}\sum_{i=1}^{n_{r_h}}(y_{r_h i} - \bar{y}_{r_h})^2 \tag{5.13}$$

[2] Although this is the traditional notation for the weight of a *stratum* in stratified sampling (e.g., Cochran 1977:90) weighting adjustments for non-response should not be confused with stratified sampling. The sampling design remains simple random sampling throughout this chapter.

5.3.2. Generalised Non-response Adjusted Estimators

In the case of a uniform response mechanism within sub-populations, a generalised π*-unbiased estimator of the population total (see section 1.3.1.1) can be constructed as:

$$\hat{Y}_{\pi^*} = \sum_{h=1}^{H}\sum_{i=1}^{n_{r_h}} \omega_{hi} y_{r_{hi}} = \sum_{h=1}^{H}\sum_{i=1}^{n_{r_h}} \frac{y_{r_{hi}}}{\varphi_{hi}\pi_{hi}} \tag{5.14}$$

where $\omega_{hi} = \dfrac{1}{\pi_{hi}\varphi_{hi}}$.

Under the uniform response mechanism within sub-populations, it is assumed that $\varphi_{hi} = \varphi_h > 0$ for all $i = 1, \ldots, n_h$ and $h = 1, \ldots, H$. The value of φ_h under the model can be obtained as (Lessler & Kalsbeek 1992:183):

$$\varphi_h = \frac{\sum_{i=1}^{n_{r_h}} 1/\pi_{hi}}{\sum_{i=1}^{n_h} 1/\pi_{hi}} \tag{5.15}$$

Equation (5.15) is an unbiased *estimator* of the mean response probability in the sub-population h, namely $\overline{\varphi}_h = \dfrac{1}{N_h}\sum_{j=1}^{N_h}\varphi_{h_j}$.

The generalised non-response adjusted π*-estimator of the population mean is:

$$\hat{\overline{Y}}_{\pi^*} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_{r_h}} \omega_{hi} y_{r_{hi}}}{\sum_{h=1}^{H}\sum_{i=1}^{n_{r_h}} \omega_{hi}} \tag{5.16}$$

5.3.3. Sample Weighting Adjustments

5.3.3.1. Non-response Adjusted Estimator of the Population Mean Assuming srs wor

When simple random sampling or any other epsem design is used to select the sample, $\pi_{hi} = \dfrac{n}{N}$ for all $i = 1, \ldots, n_h$ and all $h = 1, \ldots, H$. Conditional on the realised sample, all elements in the h-th weighting class have the same response probability of

$\varphi_h = \dfrac{n_{r_h}}{n_h}$ from (5.15), so that $\omega_{hi} = \dfrac{N}{n} \times \dfrac{n_h}{n_{r_h}}$ for $i = 1, \ldots, n_{r_h}$ and $h = 1, \ldots, H$.

From (5.16) the sample weighted estimator of $\bar{Y}$ is:

$$\hat{\bar{Y}}_{sw} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_{r_h}} \frac{n_h}{n_{r_h}} y_{r_h i}}{\sum_{h=1}^{H}\sum_{i=1}^{n_{r_h}} \frac{n_h}{n_{r_h}}} = \frac{\sum_{h=1}^{H} n_h \bar{y}_{r_h}}{\sum_{h=1}^{H} n_h} = \sum_{h=1}^{H} w_h \bar{y}_{r_h} \tag{5.17}$$

The expected value of $\hat{\bar{Y}}_{sw}$, conditional on $n_{r_h} > 0$ for all classes, is (Kalton 1983a:50):

$$E(\hat{\bar{Y}}_{sw} \mid n_{r_h}) = \sum_{h=1}^{H} w_h \bar{Y}_{r_h} \tag{5.18}$$

From (5.18) the conditional bias of the sample weighted estimator is (Little 1986:144):

$$Bias(\hat{\bar{Y}}_{sw} \mid n_{r_h}) = \sum_{h=1}^{H} (w_h - W_h)\bar{Y}_{r_h} + \sum_{h=1}^{H} W_h\left(\bar{Y}_{r_h} - \bar{Y}_h\right) \tag{5.19}$$

If $E(w_h) = W_h$, the bias of the sample weighted estimator is:

$$\begin{aligned} Bias(\hat{\bar{Y}}_{sw}) &= \sum W_h \bar{Y}_{r_h} - \bar{Y} \\ &= \sum W_h \tilde{R}_h (\bar{Y}_{r_h} - \bar{Y}_{nr_h}) \end{aligned} \tag{5.20}$$

Under the model of data missing at random within weighting classes, $E(\bar{Y}_{r_h}) = E(\bar{Y}_{nr_h})$ and $\hat{\bar{Y}}_{sw}$ is an approximately unbiased estimator of $\bar{Y}$.

The conditional variance of the estimator $\hat{\bar{Y}}_{sw}$ is (Little 1986:144):

$$V(\hat{\bar{Y}}_{sw} \mid n_{r_h}) = \sum_{h=1}^{H} w_h^2 \left(1 - \frac{n_{r_h}}{N_h}\right) \frac{S_{r_h}^2}{n_{r_h}} \tag{5.21}$$

Subject to the requirement that $n_{r_h} \geq 2$ for all $h = 1, \ldots, H$, $S_{r_h}^2$ in (5.21) may be estimated by $s_{r_h}^2$. However, if some of the n_{r_h} are very small, the $s_{r_h}^2$ may be poor estimates (Oh & Scheuren 1983:151). Since the N_h are unknown, (5.21) can be estimated by (Oh & Scheuren 1983:152):

$$v(\hat{\bar{Y}}_{sw} \mid n_{r_h}) = \sum_{h=1}^{H} w_h^2 \left(1 - \frac{n n_{r_h}}{n_h N}\right) \frac{s_{r_h}^2}{n_{r_h}} \tag{5.22}$$

The approximate unconditional variance of $\hat{\bar{Y}}_{sw}$, assuming that N is large (i.e., ignoring the fpc), is given by Kalton (1983a:52) as:

$$AV(\hat{\bar{Y}}_{sw}) = \sum_{h=1}^{H} \frac{W_h S_{r_h}^2}{R_h n} + \sum_{h=1}^{H} \frac{\tilde{R}_h S_{r_h}^2}{R_h^2 n^2} + \sum_{h=1}^{H} \frac{1}{n} W_h (\bar{Y}_{r_h} - \hat{\bar{Y}}_{sw})^2 \tag{5.23}$$

For large samples, the approximate unconditional variance may be estimated by (Kalton 1983a:54):

$$v(\hat{\bar{Y}}_{sw}) = \sum \frac{w_h^2 s_{r_h}^2}{n_{r_h}} + \frac{\sum w_h (\bar{y}_h - \hat{\bar{Y}}_{sw})^2}{n} \tag{5.24}$$

5.3.3.2. Non-response Adjusted Estimator of the Population Total Assuming srs wor

From (5.14) the sample weighted estimator of the population total is (Oh & Scheuren 1983:150):

$$\hat{Y}_{sw} = \sum_{h=1}^{H} \sum_{i=1}^{n_{r_h}} \omega_{hi} y_{r_{hi}} = \frac{N}{n} \sum_{h=1}^{H} n_h \bar{y}_{r_h} = N \sum_{h=1}^{H} w_h \bar{y}_{r_h} \tag{5.25}$$

The conditional bias of the estimator $\hat{Y}_{sw}$ is (Lehtonen & Pahkinen 1995:122):

$$Bias(\hat{Y}_{sw} \mid n_{r_h}) = N \sum_{h=1}^{H} (w_h - W_h) \bar{Y}_{r_h} + N \sum_{h=1}^{H} W_h \left(\bar{Y}_{r_h} - \bar{Y}_h\right) \tag{5.26}$$

If $E(w_h) = W_h$, the bias of the sample weighted estimator is:

$$Bias(\hat{Y}_{sw}) = N \sum_{h=1}^{H} W_h \left(\bar{Y}_{r_h} - \bar{Y}_h\right) \tag{5.27}$$

Under the model of data missing at random within the weighting classes, $E(\bar{Y}_{r_h}) = E(\bar{Y}_{nr_h})$ and $\hat{Y}_{sw}$ is an approximately unbiased estimator of Y.

The approximate unconditional variance of $\hat{Y}_{sw}$ is (Oh & Scheuren 1983:154):

$$AV(\hat{Y}_{sw}) = \frac{N}{n}\left(\frac{N-n}{N-1}\right)\sum_{h=1}^{H} N_h\left(\bar{Y}_h - \bar{Y}\right)^2 + \sum_{h=1}^{H} N_h^2\left(1-\frac{\bar{n}_{r_h}}{N_h}\right)\frac{S_h^2}{\bar{n}_{r_h}} + \sum_{h=1}^{H} N_h^2\left(1-\frac{\bar{n}_{r_h}}{\bar{n}_h}\right)\frac{S_h^2}{\bar{n}_{r_h}^2} \tag{5.28}$$

where $\bar{n}_h = n\left(\frac{N_h}{N}\right)$ and $\bar{n}_{r_h} = n\left(\frac{N_h}{N}\right)\varphi_h$.

5.3.4. Population Weighting Adjustments

In the case of population weighting adjustments, the respondent data in each weighting class are weighted by a constant factor so that the weighted distribution over the classes conforms to the population distribution. The population distribution is presumed known without error.

5.3.4.1. Non-response Adjusted Estimator of the Population Mean Assuming srs wor

In the case of sample weighting adjustments, the estimator of the population mean was shown to be $\hat{\bar{Y}}_{sw} = \sum w_h \bar{y}_{r_h}$, where $w_h = \frac{n_h}{n}$, but when the population proportions $W_h = \frac{N_h}{N}$ are known, the estimator of $\bar{Y}$ using population weighting adjustment is (Kalton 1983a:50):

$$\hat{\bar{Y}}_{pw} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_{r_h} = \sum_{h=1}^{H} W_h \bar{y}_{r_h} \tag{5.29}$$

Noting that the overall population mean can be written as:

$$\begin{aligned}\bar{Y} &= \frac{1}{N}\sum_{h=1}^{H} N_h \bar{Y}_h = \sum W_h \bar{Y}_h \\ &= \sum W_h \bar{Y}_{r_h} - \sum W_h \tilde{R}_h (\bar{Y}_{r_h} - \bar{Y}_{nr_h})\end{aligned} \tag{5.30}$$

the conditional bias of $\hat{\bar{Y}}_{pw}$ (given that $n_{r_h} > 0$ for all h) can be written as (Kalton 1983a:45):

$$\begin{aligned}Bias(\hat{\bar{Y}}_{pw} \mid n_{r_h}) &= \sum W_h \bar{Y}_{r_h} - \bar{Y} \\ &= \sum W_h \tilde{R}_h (\bar{Y}_{r_h} - \bar{Y}_{nr_h})\end{aligned} \tag{5.31}$$

Clearly, the estimator $\hat{\bar{Y}}_{pw}$ will be approximately unbiased under the model of data missing at random within each weighting class.

If the expected number of respondents is large in all classes, the conditional variance of the population weighted estimator of the mean is (Little 1986:144):

$$V(\hat{\bar{Y}}_{pw} \mid n_{r_h}) = \sum W_h^{\,2}\left(1 - \frac{n_{r_h}}{N_h}\right)\frac{S_{r_h}^{\;2}}{n_{r_h}} \tag{5.32}$$

An unbiased estimator of the conditional variance is:

$$v(\hat{\bar{Y}}_{pw} \mid n_{r_h}) = \sum W_h^{\,2}(1 - \frac{n_{r_h}}{N_h})\frac{s_{r_h}^{\;2}}{n_{r_h}} \tag{5.33}$$

5.3.4.2.Non-response Adjusted Estimator of the Population Total Assuming srs wor

The population weighted estimator of the population *total* is (Oh & Scheuren 1983:150):

$$\hat{Y}_{pw} = \sum_{h=1}^{H} N_h \bar{y}_{r_h} \tag{5.34}$$

The conditional bias of this estimator is (Lehtonen & Pahkinen 1995:122):

$$Bias(\hat{Y}_{pw} \mid n_{r_h}) = N\sum_{h=1}^{H} W_h\left(\bar{Y}_{r_h} - \bar{Y}_h\right) \tag{5.35}$$

which can be written in a manner similar to (5.31) as:

$$Bias(\hat{Y}_{pw} \mid n_{r_h}) = \sum N_h \tilde{R}_h(\bar{Y}_{r_h} - \bar{Y}_{nr_h}) \tag{5.36}$$

The conditional variance of $\hat{Y}_{pw}$ is (Lehtonen & Pahkinen 1995:122):

$$V(\hat{Y}_{pw} \mid n_{r_h}) = \sum N_h^{\,2}\left(1 - \frac{n_{r_h}}{N_h}\right)\frac{S_{r_h}^{\;2}}{n_{r_h}} \tag{5.37}$$

5.3.4.3.Resemblance to the Post-stratified Estimator

Although population weighting adjustment bears a close resemblance to post-stratification, there are a number of important differences between the techniques:

1. Both techniques weight the sample to make the sample distribution conform to the population distribution across a set of strata or classes, but where post-stratification is exclusively concerned with random sampling variability in the spread of the sample, population weighting adjustment also (but *primarily*) corrects for the variation in response rates (Kalton 1983a:45).

2. In the case of post-stratification, the stratum means ($\bar{y}_h$) are unbiased estimators of the population stratum means ($\bar{Y}_h$); in the case of population weighting adjustments, the class sample means ($\bar{y}_{r_h}$) are unbiased estimators of the *respondent* class means ($\bar{Y}_{r_h}$) and will be unbiased estimators of the overall class means only if the response mechanism has been modelled correctly (Kalton 1983a:45).

3. In most cases, post-stratification involves relatively minor weighting adjustments which may be thought of as fine-tuning (Kalton & Kasprzyk 1986:2). On the other hand, population weighting adjustments often involve more major adjustments which may lead to increases in variances.

4. The variance of the post-stratified estimator from a simple random sample with large expected stratum sizes, cannot exceed the variance of the mean from a simple random sample of the same size. Under similar conditions, however, the variance of the population weighting adjustment may exceed $V(\bar{y}_r)$ as will be shown in section 5.3.6 (Kalton 1983a:47).

Some general remarks regarding the formation of weighting classes, applicable to both population weighting adjustments and sample weighting adjustments, are made in section 5.4.

5.3.5. Effect of Weighting Adjustments on Non-response Bias

To illustrate the effect of weighting adjustments on non-response bias, consider the unweighted respondent mean $\bar{y}_r$ from a simple random sample selected from a population which has been divided into weighting classes. The bias of this estimator may be expressed as the sum of two components (Kalton 1983a:45);

$$Bias(\bar{y}_r) = \frac{\sum W_h(\bar{Y}_{r_h} - \bar{Y}_r)(R_h - R)}{R} + \sum W_h \tilde{R}_h(\bar{Y}_{r_h} - \bar{Y}_{nr_h}) \tag{5.38}$$
$$= A + B$$

Component A arises from the variability in the response rates among the classes and component B arises from the differences between respondent and non-respondent means within classes. From a comparison of (5.19) and (5.38) can be seen that the bias of the sample weighted estimator is:

$$Bias(\hat{\bar{Y}}_{sw}) = \sum_{h=1}^{H}(w_h - W_h)\bar{Y}_{r_h} + B \tag{5.39}$$

but if $E(w_h) = W_h$, then $Bias(\hat{\bar{Y}}_{sw}) = B$.

From (5.31) can be seen that the bias of the population weighted sample mean is:

$$Bias(\hat{\bar{Y}}_{pw}) = \sum W_h \tilde{R}_h(\bar{Y}_{r_h} - \bar{Y}_{nr_h}) = B \tag{5.40}$$

Hence, if $E(w_h) = W_h$, then:

$$Bias(\hat{\bar{Y}}_{sw}) = Bias(\hat{\bar{Y}}_{pw}) = B \tag{5.41}$$

From a comparison of (5.38) and (5.41) various conclusions can be made:

1. Weighting adjustments will reduce or (ideally) eliminate the non-response bias ($B \approx 0$) if the assumption of data missing at random within weighting classes is correct, i.e., if $E(\bar{Y}_{r_h}) = E(\bar{Y}_{nr_h})$. This condition illustrates the importance of forming suitable weighting classes (see section 5.4). In practice, however, weighting classes are usually not completely homogeneous so that some bias remains after weighting adjustments are applied.

2. From (5.38) and (5.41) can be seen that:

   a) If components A and B have the same sign, the absolute bias will be reduced by $|A|$.

   b) If components A and B have opposite signs, the bias will be reduced if and only if $2|B| < |A|$ (Thomsen 1973:279).

From (5.38) can be seen that $|A|$ is large when the response rates R_h vary considerably among large classes[3] or when there are large differences among the respondent means in these classes. Under these conditions, weighting adjustments will be very effective in reducing non-response bias.

3. Weighting adjustments have no effect on non-response bias if component A is zero, i.e., if:

   a) the response rates do not vary among the classes ($R_h = R$ for all h) or

   b) the respondent means do not vary among the classes ($\bar{Y}_{r_h} = \bar{Y}_r$ for all h).

4. Weighting adjustments may *increase* the absolute bias of the estimate if A and B have opposite signs and $|A| \leq 2|B|$.

The above four conclusions will be illustrated in section 5.5 by means of three hypothetical examples.

5.3.6. Effect of Weighting Adjustments on the Variance

The approximate conditional variance of the unweighted estimator $\bar{y}_r$ (ignoring the fpc) is expressed by Kalton (1983a:47) as:

$$AV(\bar{y}_r \mid n_{r_h}) = \frac{1}{n_r}\left[\sum R_h\left(\bar{Y}_{r_h} - \bar{Y}_r\right)^2 + \sum R_h S_{r_h}{}^2\right] \tag{5.42}$$

A comparison between the conditional variance of the population weighted estimator (5.32) with (5.42) reveals that no general conclusions regarding the precision of the population weighted estimator $\hat{\bar{Y}}_{pw}$ versus that of the unweighted estimator $\bar{y}_r$ can be made. Some conclusions can, however, be made in special cases (Kalton 1983a:48):

1. If $R_h = R$ for all $h = 1, \ldots, H$, i.e., if the response rate is constant across classes, $V(\hat{\bar{Y}}_{pw} \mid n_{r_h}) < AV(\bar{y}_r \mid n_{r_h})$. This suggests a bias-variance trade-off in the formation of weighting classes since large differences among the response rates

[3] Large variability in the R_h through many (smaller) classes will not greatly affect component A, because of their smaller contributions through W_h.

are desirable for a reduction in bias but, on the other hand, lead to an increase in the variance.

2. If $S_{r_h}{}^2 = S_r{}^2$, $R_h \neq R$ and $\bar{Y}_{r_h} \approx \bar{Y}_r$ for all h = 1, ..., H, i.e., if the element variances within the classes are equal and the classes differ substantially in their response rates but only slightly in their means, $V(\hat{\bar{Y}}_{pw} \mid n_{r_h}) > AV(\bar{y}_r \mid n_{r_h})$. This result shows that population weighting adjustment is harmful to precision when weighting classes are formed with equal class means.

Kalton (1983a:153) compares the approximate unconditional variance of the sample weighted estimator $\hat{\bar{Y}}_{sw}$ (5.23) with the approximate variance of $\bar{y}_r$:

$$AV(\bar{y}_r) = \frac{1}{nR}\left[\sum R_h(\bar{Y}_{r_h} - \bar{Y}_r)^2 + \sum R_h S_{r_h}{}^2\right] \tag{5.43}$$

Although no general conclusions on the sign or magnitude of the difference between the two variances can be made, Kalton (1983a:53) considers the special case where $S_{r_h}{}^2 = S_r{}^2$ and the variances among the class means $\bar{Y}_{r_h}$ are negligible compared to the within-class variances. In this case, the use of variable weights in $\hat{\bar{Y}}_{sw}$ leads to the result that $V(\hat{\bar{Y}}_{sw}) \geq AV(\bar{y}_r)$.

Under the above conditions, weighting increases the variance of a sample mean by an approximate factor (Kalton & Kasprzyk 1986:4):

$$L = \frac{\sum n_h\left(\sum n_h \omega_h{}^2\right)}{\left(\sum n_h \omega_h\right)^2} \tag{5.44}$$

The factor L becomes large when there is a large variation in the weights ω_h. A large variation in weights may arise from segmenting the sample into many weighting classes with small sample sizes in each, leading to unstable class response rates. To avoid this effect, it is common practice to limit the number of weighting classes. If there are still some weighting classes that require large weights, these weighting classes are handled by either collapsing them with adjacent ones or cutting back their weights to some acceptable maximum value (a maximum weight of 2 has been mentioned). Although limiting the weights or collapsing weighting classes may restrict the increase in variance associated with the use of extreme weights, they may lead to increased bias (see section 5.4); their effect on the bias is, however, unknown. (Kalton & Kasprzyk 1986:4.)

To compare the variances of the population weighted estimator and the sample weighted estimator of the population mean, Kalton (1983a:53) derives the approximate unconditional variance of $\hat{\bar{Y}}_{pw}$ as:

$$AV(\hat{\bar{Y}}_{pw}) = \sum_{h=1}^{H} \frac{W_h S_h^2}{R_h n} \tag{5.45}$$

which is the first term on the right-hand side of (5.23) (replacing S_h^2 with its unbiased estimator $S_{r_h}^2$). To the order of approximation used, the second term on the right-hand side of (5.23) is negligible, so that $AV(\hat{\bar{Y}}_{sw})$ exceeds $AV(\hat{\bar{Y}}_{pw})$ by the non-negative third term. The variance of the sample weighted estimator may be considerably larger than that of the population weighted estimator unless the n_h are reasonably large. Since $Bias(\hat{\bar{Y}}_{pw}) = Bias(\hat{\bar{Y}}_{sw})$ if $E(w_h) = W_h$, but $AV(\hat{\bar{Y}}_{sw}) \geq AV(\hat{\bar{Y}}_{pw})$, it follows that $MSE(\hat{\bar{Y}}_{sw}) \geq MSE(\hat{\bar{Y}}_{pw})$. (Kalton 1983a:53.)

In practice, population and sample weighting adjustments are often used in combination. In fact, even when sample weighting adjustments are being used, it is recommended that population weighting adjustments are employed as well in order to deal with under-coverage (Kalton 1983a:44). According to Kalton and Kasprzyk (1986:4) a general procedure is to first determine the sample weights to compensate for unequal inclusion probabilities, then to revise these weights to compensate for unequal response rates in different weighting classes and finally to revise these weights to make the weighted sample distribution for certain characteristics conform to the known population distribution for those characteristics.

5.4. FORMATION OF WEIGHTING CLASSES

Various authors, for example, Oh and Scheuren (1983), Kalton (1983a), Kalton and Kasprzyk (1986) and Little (1986) give suggestions regarding the formation of weighting classes. Nevertheless, in large-scale surveys with many diverse survey variables, it may be impossible to attain an overall feasible construction for the weighting classes (Lehtonen & Pahkinen 1995:123).

Two key dimensions in the choice of weighting classes may be distinguished, namely response probabilities and the values of population means among classes. Weighting classes are sometimes determined by the use of CHAID analysis (see section 3.2.2).

5.4.1. Homogeneity of Weighting Classes

The statistical properties of the population weighted estimator and the sample weighted estimator depend on the formation of weighting classes that are internally homogeneous and externally heterogeneous with respect to the survey variable so that the assumption of a uniform response probability within the classes is tenable. The auxiliary variables used to define the weighting classes should therefore be highly correlated with the major survey variables but should be mutually unrelated.

In a multi-purpose survey where a number of survey variables are being observed, it is quite possible that the non-response will affect survey variables differently. For some, the only effect of non-response may be to reduce the sample size, i.e., there may be no bias impact whatsoever. For other survey variables, there may be some bias but the non-response adjustment may reduce or essentially eliminate it. For the remainder of the variables, the adjustment may leave residual biases or could conceivably even increase their bias.

5.4.2. Number of Weighting Classes

In deciding on the number H of weighting classes to use, one is torn between two competing influences. On the one hand, a large number of well-formed classes will more effectively reduce the bias but the wider variation in weights caused by more variable weighting class adjustments will cause the variance of estimates to increase.

Thus, keeping H small is likely to reduce the variance but increasing H is likely to reduce the bias. (Lessler & Kalsbeek 1992:188.)

Oh and Scheuren (1983:155) suggest that, if the means of two sub-populations are suspected of being equal or nearly so, then the classes may be pooled to attempt to reduce the variance increase. A difficulty with this prescription is, however, that the survey variable mean can only be observed for the respondents. Rules to pool or collapse smaller classes, although reducing the degree to which bias may be eliminated, do have the effect of constraining the variance increase that the adjustment may create.

5.4.3. Size of Weighting Classes

All weighting classes should be large enough so that the n_{r_h} for $h = 1, \ldots, H$ are sufficiently great (say 20 or more). Adjacent classes may be pooled when one or more of the n_{r_h} is small. Oh and Scheuren (1983:161) state:

> *... the requirement of a robust adjustment for non-response would seem to necessitate fairly high response rates. Without good response our estimators depend far too heavily on specific knowledge of the underlying mechanism - information we rarely have in practice.*

5.4.4. Accuracy of Population Distribution

When performing population weighting adjustments, the measures of N_h should be reasonably accurate. Although perfect measures are seldom available, the measures used for N_h should preferably be of higher quality than estimates of N_h that could be obtained directly from the survey data.

5.4.5. Number of Auxiliary Variables

In the case where one has several suitable auxiliary variables from which to choose, it is usually preferable to form classes by crossing a few levels on all variables than to pick the best one and divide it into numerous levels. It also seems preferable to define classes by a coarse division on several acceptable variables than to form the same number of classes by a finer division on the best single variable among those considered acceptable. (Lessler & Kalsbeek 1992:188.)

5.5. THREE SIMPLE EXAMPLES

5.5.1. Example 1

Suppose a survey is conducted to determine the mean monthly expenditure on water and electricity of households in a certain district in the Eastern Cape. Suppose an epsem sample of size 1000 is selected from the population of 10000 households and the final response rate is 67,9%.

The unadjusted mean of the 679 respondents is $\bar{y}_r$ = R89,50. Suppose the population mean (to be estimated) is $\bar{Y}$ = R81,45. The bias of the respondent mean is from (5.38):

$$Bias(\bar{y}_r) = R89{,}50 - R81{,}45 = R8{,}06 = A + B$$

The unadjusted estimator overestimates the mean expenditure on water and electricity by R8,06. Component A (which arises from the variability in response rates among sub-populations) and component B (which arises from differences among the respondent and non-respondent means within sub-populations) are, at this stage, unknown.

In order to perform non-response adjustments, two auxiliary variables are identified which are presumed to be correlated with household expenditure as well as with response probability. These auxiliary variables are race and urbanity. Interviewers are instructed to obtain the race of non-respondents from e.g., neighbours while the urbanity of each sample element is known from the sampling frame. The population is assumed to consist of the four sub-populations formed by a cross-classification of race ("white" and "black") and urbanity ("urban" and "rural"). Suppose the unknown values of the sub-population means are as given in Table 5.1. Table 5.1 also gives the population and sample sizes, the number of respondents, the response rate and the respondent mean for each of the four sub-populations.

Table 5.1 Sub-population values for Example 1

	Urban-Black	Urban-White	Rural-White	Rural-Black
N_h	4600	1700	950	2750
n_h	500	150	80	270
n_{r_h}	400	135	36	108
r_h	0,80	0,90	0,45	0,40
$\bar{y}_{r_h}$	R86	R150	R110	R20
$\bar{Y}_h$	R90	R152	R109	R14

The effectiveness (in terms of bias reduction) of sample and population weighting will now be examined by comparing the biases of the sample and population weighted estimates with the bias of the unadjusted mean $\bar{y}_r$.

a) Sample weighted estimate

Suppose the sub-population sizes N_h are unknown. The sample weighted estimate is from (5.17):

$$\begin{aligned}\hat{\bar{Y}}_{sw} &= \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{r_h} \\ &= \frac{500}{1000}(86) + \frac{150}{1000}(150) + \frac{80}{1000}(110) + \frac{270}{1000}(20) \\ &= R79{,}70\end{aligned}$$

The bias of this estimate is:

$$Bias(\hat{\bar{Y}}_{sw}) = R79{,}70 - R81{,}45 = -R1{,}75 \approx B$$

from (5.41). Use of sample weighting has resulted in a considerable bias reduction in the estimate of the population mean relative to that of the unadjusted mean.

b) Population weighted estimate

Suppose the sub-population sizes N_h are known. The population weighted estimate is from (5.29):

$$
\begin{aligned}
\hat{\bar{Y}}_{pw} &= \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{r_h} \\
&= \frac{4600}{10000}(86) + \frac{1700}{10000}(150) + \frac{950}{10000}(110) + \frac{2750}{10000}(20) \\
&= R81{,}01
\end{aligned}
$$

The bias of the population weighted estimate is from (5.40):

$$Bias(\hat{\bar{Y}}_{pw}) = R81{,}01 - R81{,}45 = -R0{,}44 = B$$

Population weighting has been slightly more effective in bias reduction than sample weighting. The difference in bias reduction between the two estimates is because of differences between the sample proportions w_h and the population proportions W_h.

Component B of the bias would have been zero, i.e., population weighting would have completely eliminated non-response bias if the weighting classes were defined in such a way that there are no differences between respondents and non-respondents within the weighting classes. In this example, there seem to be relatively small differences between respondent and non-respondent means within the four classes ($\bar{y}_{r_h}$ does not differ much from $\bar{Y}_h$), resulting in a relatively small residual non-response bias.

Since component B of the bias is equal to $-R0{,}44$, component A is equal to $A = R8{,}06 - (-R0{,}44) = R8{,}50$. Note that components A and B have opposite signs. In this case, the requirement for bias reduction (see section 5.3.5) is satisfied, i.e., $2|B| = 2|-0{,}44| < |9{,}81|$.

The bias reduction due to weighting adjustment is relatively large in this example due to the considerable variation in the response rates among the classes. There would have been no reduction in bias if the response rates did not vary among the classes and the respondent means did not vary among the classes.

5.5.2. Example 2

With reference to the survey in Example 1, suppose the survey variable is the mean annual income of households in the Eastern Cape district. The same weighting classes as in Example 1 are used to adjust for non-response. Suppose the unknown values of the sub-population means are as given in Table 5.2 and the overall population mean (to be estimated) is $\bar{Y}$ = R18 217.

Table 5.2 Sub-population values for Example 2

	Urban-Black	Urban-White	Rural-White	Rural-Black
N_h	4600	1700	950	2750
n_h	500	150	80	270
n_{r_h}	400	135	36	108
r_h	80%	90%	45%	40%
$\bar{y}_{r_h}$	R11 160	R35 880	R25 460	R7 128
$\bar{Y}_h$	R14 350	R42 760	R26 300	R6 720

a) Unadjusted estimate

The overall respondent mean is $\bar{y}_r$ = R16 192. The bias of this estimate is from (5.38):

$$Bias(\bar{y}_r) = R16192 - R18217 = -R2025 = A + B$$

b) Sample weighted estimate

Suppose the population sizes N_h are unknown. The sample weighted estimate is from (5.17):

$$\begin{aligned}\hat{\bar{Y}}_{sw} &= \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{r_h} \\ &= \frac{500}{1000}(11160) + \frac{150}{1000}(35880) + \frac{80}{1000}(25460) + \frac{270}{1000}(7128) \\ &= R14923\end{aligned}$$

The bias of this estimate is from (5.41):

$$Bias(\hat{\bar{Y}}_{sw}) = R14923 - R18217 = -R3294 \approx B$$

In this example, sample weighting has actually led to an *increase* in non-response bias relative to the unadjusted estimator $\bar{y}_r$.

c) Population weighted estimate

Suppose the population sizes are known. The population weighted estimate is from (5.29):

$$
\begin{aligned}
\hat{\bar{Y}}_{pw} &= \sum \frac{N_h}{N} \bar{y}_{r_h} \\
&= \frac{4600}{10000}(11160) + \frac{1700}{10000}(35880) + \frac{950}{10000}(25460) + \frac{2750}{10000}(7128) \\
&= R15612
\end{aligned}
$$

The bias of the population weighted estimate is from (5.40):

$$Bias(\hat{\bar{Y}}_{pw}) = -R2605 = B$$

Population weighting has also led to an increase in non-response bias relative to the unadjusted estimate but the increase is less severe than in the case of sample weighting. The reasons for the failure of population and sample weighting to reduce non-response bias in this example may be found when examining the components A and B of the bias: For the population weighted estimate, component A is $A = -R2025 - (-R2605)$ $= R1269$. In this example, components A and B have opposite signs and $2|2605| \nless |1269|$ so that the bias is increased. Similarly for the sample weighted estimate.

The relatively large component B, i.e., the residual bias after weighting adjustments, arises from the differences between respondent and non-respondent means within classes. This component would have been small, i.e., weighting adjustments would have been effective if the mean household income for respondents and non-respondents in each of the sub-populations were approximately equal. In this example, the mean income for non-respondents differs substantially from the mean income for respondents within each class. Since household income is usually under-estimated by respondent data, component B is relatively large and negative.

This example illustrates the importance of forming homogeneous weighting classes. To the extent that weighting classes formed for non-response adjustment are *homogeneous*, i.e., little variation in response rates within the classes and similarity between respondents and non-respondents within classes, weighting adjustments will be successful in reducing the bias.

5.5.3. Example 3

With reference to the survey in Example 2, suppose weighting classes are based on an auxiliary variable which is not strongly correlated with response behaviour, say geographic area: northern, southern, eastern and western part of the Eastern Cape. Suppose the unknown values of the sub-population means are as given in Table 5.3 and the overall population mean (to be estimated) is, as in Example 2, equal to $\bar{Y}$ = R18 217.

Table 5.3 Sub-population values for Example 3

	North	South	East	West
N_h	4600	2750	1700	950
n_h	500	270	150	80
n_{r_h}	347	176	102	54
r_h	69%	65%	68%	68%
$\bar{y}_{r_h}$	R16099	R15900	R17060	R16100
$\bar{Y}_h$	R17113	R18 900	R19 540	R19 217

a) Unadjusted estimate

The overall respondent mean is $\bar{y}_r$ = R16 192 and the bias of this estimate is as before:

$$Bias(\bar{y}_r) = R16192 - R18217 = -R2025 = A + B$$

b) Sample weighted estimate

Suppose the population sizes N_h are unknown. The sample weighted estimate is from (5.17):

$$
\begin{aligned}
\hat{\bar{Y}}_{sw} &= \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{r_h} \\
&= \frac{500}{1000}(16099) + \frac{270}{1000}(15900) + \frac{150}{1000}(17060) + \frac{80}{1000}(16100) \\
&= R16189{,}50
\end{aligned}
$$

The bias of this estimate is from (5.41):

$$
\begin{aligned}
Bias(\hat{\bar{Y}}_{sw}) &= R16189{,}50 - R18217 = -R2027{,}5 \\
&\approx Bias(\bar{y}_r) = -R2025
\end{aligned}
$$

Sample weighting has not led to any significant change in the bias of the unadjusted estimator due to the "bad choice" of auxiliary variable.

c) Population weighted estimate

Suppose the population sizes are known. The population weighted estimate is from (5.29):

$$
\begin{aligned}
\hat{\bar{Y}}_{pw} &= \sum \frac{N_h}{N} \bar{y}_{r_h} \\
&= \frac{4600}{10000}(16099) + \frac{1700}{10000}(15900) + \frac{950}{10000}(17060) + \frac{2750}{10000}(16100) \\
&= R16157
\end{aligned}
$$

The bias of the population weighted estimate is from (5.40):

$$
\begin{aligned}
Bias(\hat{\bar{Y}}_{pw}) &= R16157 - R18217 = -R2060 \\
&\approx Bias(\bar{y}_r) = -R2025
\end{aligned}
$$

In this example $A = -R2025 - (-R2060) = R35$. This is the component due to the variability in response rates among classes which, in this case, is relatively small. This example illustrates that weighting adjustments for non-response have very little effect on non-response bias if weighting classes are formed in such a way that response rates do not vary among the classes and/or the respondent means do not differ among the classes.

5.6. RAKING RATIO ESTIMATION

Raking ratio estimation, or briefly "raking", also known as *iterative proportional fitting* (IPF) in contingency table analysis (Bishop, Fienberg & Holland 1975:Chapter 3), entails the estimation of cell frequencies in an $H \times K$ contingency table for which the marginal totals are known[8]. Raking was first proposed in the US Population Census of 1940 as a method for obtaining conformity between population and sample "counts". Raking has since been used for a wide variety of problems including that of adjusting for non-response. (Oh & Scheuren 1983:162.) In section 5.6.1, raking is first discussed in the context of surveys where the population marginals of the cross-classified survey variables are known and the aim of the raking procedure is to obtain estimates of the population *cell* frequencies - the problem of non-response is ignored. In section 5.6.2 raking is discussed when the aim is to adjust for non-response. In this case, a number of *adjustment cells* are formed by cross-classifying certain auxiliary variables whose population marginals are known. The values of the survey variables are available only for a random sample from this population.

5.6.1. The General Raking Technique

5.6.1.1. The Iterative Adjustment Process

Suppose a sample of size n is selected from a population of N elements. The sample elements are cross-classified into a 2-dimensional contingency table in terms of two survey variables whose row and column marginals are known for the entire population[9]. Suppose one survey variable has H categories and the other has K categories, giving HK cells, called *adjustment* cells. The notation used to denote population and sample joint and marginal frequencies for these two variables are set out in Table 5.4 and Table 5.5. The population row and column marginals ($N_{h.}$ and $N_{.k}$) are assumed to be known but the cell frequencies N_{hk} are unknown. Clearly, the sample frequencies, n_{hk} are known.

[8] The term "raking" for this procedure seems to derive from the analogue between the use of an ordinary garden rake and the successive steps in the iterative proportional fitting method. According to Oh and Scheuren (1983:163) to set this analogy requires a good bit of experience both in gardening and in doing the iterations by hand!

[9] There is theoretically no limit to the number of dimensions that could be handled but, in practice, a limit has to be imposed on the number of dimensions for the sake of convergence.

Table 5.4 Population joint and marginal frequencies

	1	2	...	K	Total
1	N_{11}	N_{12}	...	N_{1K}	$N_{1\cdot}$
2	N_{21}	N_{22}	...	N_{2K}	$N_{2\cdot}$
...	...	...	...	...	...
H	N_{H1}	N_{H2}	...	N_{HK}	$N_{H\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	...	$N_{\cdot K}$	N

Table 5.5 Sample joint and marginal frequencies

	1	2	...	K	Total
1	n_{11}	n_{12}	...	n_{1K}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2K}	$n_{2\cdot}$
...	...	...	...	...	...
H	n_{H1}	n_{H2}	...	n_{HK}	$n_{H\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot K}$	n

The objective of raking is to obtain estimates, say, $\hat{N}_{hk}$ of N_{hk} such that the following two equations (called "constraint equations") are simultaneously satisfied (Oh & Scheuren 1978:716):

$$\hat{N}_{h\cdot} = \sum_{k=1}^{K} \hat{N}_{hk} = N_{h\cdot} \tag{5.46}$$

and

$$\hat{N}_{\cdot k} = \sum_{h=1}^{H} \hat{N}_{hk} = N_{\cdot k} \tag{5.47}$$

To derive the $\hat{N}_{hk}$, the raking algorithm proceeds by proportionately scaling the sample cell frequencies n_{hk} so that each of the constraint equations is satisfied in turn. Each step begins with the results of the previous steps and the procedure ends when the raking algorithm converges (called convergent raking ratio estimation) or when the constraints are simultaneously satisfied to the desired degree of closeness (called limited raking ratio estimation) (Oh & Scheuren 1978b:717).

For example, in the case of two survey variables with H and K categories respectively, the first cycle of the algorithm consists of the following two steps:

Step 1. Obtain the row adjusted cell frequencies:

$$n_{hk}^{(1)} = \frac{N_{h\cdot}}{n_{h\cdot}} n_{hk} = a_h^{(1)} n_{hk} \tag{5.48}$$

where $a_h^{(1)}$ is the row adjustment factor in cycle 1. The weighted row marginal frequencies $n_{h\cdot}^{(1)}$ will now be equal to the population row marginal frequencies $N_{h\cdot}$ for $h = 1, \ldots, H$, since:

$$n_{h\cdot}^{(1)} = \sum_{k=1}^{K} \frac{N_{h\cdot}}{n_{h\cdot}} n_{hk} = N_{h\cdot} \tag{5.49}$$

while:

$$n_{\cdot k}^{(1)} = \sum_{h=1}^{H} \frac{N_{h\cdot}}{n_{h\cdot}} n_{hk}^{(1)} \approx N_{\cdot k} \tag{5.50}$$

Step 2. Obtain the column adjusted frequencies, using the results of step 1:

$$\begin{aligned} n_{hk}^{(2)} &= \frac{N_{\cdot k}}{n_{\cdot k}^{(1)}} n_{hk}^{(1)} = b_k^{(1)} n_{hk}^{(1)} \\ &= a_h^{(1)} b_k^{(1)} n_{hk} = \omega_{hk}^{(2)} n_{hk} \end{aligned} \tag{5.51}$$

where $b_k^{(1)}$ is the column adjustment factor in cycle 1. The new column marginal frequencies $n_{\cdot k}^{(2)}$ will now be equal to the population column marginal frequencies $N_{\cdot k}$ for $k = 1, \ldots, K$, since:

$$n_{\cdot k}^{(2)} = \sum_{h=1}^{H} n_{hk}^{(2)} = \sum_{h=1}^{H} \frac{N_{\cdot k}}{n_{\cdot k}^{(1)}} n_{hk}^{(1)} = N_{\cdot k} \tag{5.52}$$

and:

$$n_{h\cdot}^{(2)} = \sum_{k=1}^{K} n_{hk}^{(2)} \tag{5.53}$$

These two steps constitute a cycle which is repeated in whole or in part until the cell frequencies converge or until the desired degree of closeness is obtained. According to Brackstone and Rao (1981:100) empirical evidence indicates that convergence to the desired row and column marginals may require only 4 to 5 iterations or less. According to Oh and Scheuren (1978b:717) "reasonable results" are usually achieved with raking in two to four cycles. Brackstone and Rao (1981:100) recommend using an even number of iterations if both survey variables are equally important.

In general, at the *end* of the *c*-th cycle, the estimates of N_{hk} are:

$$\hat{N}_{hk}^{(c)} = \omega_{hk}^{(c)} n_{hk} \tag{5.54}$$

where the weighting factors $\omega_{hk}^{(c)}$ may be derived in either of two ways:

1. Solve for $\omega_{hk}^{(c)}$ from the equivalent of equation (5.51) in cycle *c*. For example, at the end of cycle 4:

$$\omega_{hk}^{(4)} = \begin{cases} \dfrac{n_{hk}^{(8)}}{n_{hk}} & \textit{for } n_{hk} \neq 0 \\ 0 & \textit{for } n_{hk} = 0 \end{cases}$$

2. Keep track of the cumulative products over all the cycles required, i.e., at the end of the *c*-th cycle, the factors $\omega_{hk}^{(c)}$ may be expressed as:

$$\omega_{hk}^{(c)} = \prod_{i=1}^{c} a_h^{(i)} b_k^{(i)} \tag{5.55}$$

5.6.1.2.Statistical Properties of the Raking Estimators

Oh and Scheuren (1978b, 1983) discuss some statistical properties of the raking estimators of the marginals $\hat{N}_{h\cdot}$ and $\hat{N}_{\cdot k}$ and the cell frequencies $\hat{N}_{hk}$. *Firstly,* they note that the estimators of the row and column marginals $\hat{N}_{h\cdot}$ and $\hat{N}_{\cdot k}$ can be brought as close as desired to the "true" marginals $N_{h\cdot}$ and $N_{\cdot k}$, provided that the algorithm is convergent. *Secondly,* for large samples and in the case of simple random sampling, it can be shown that $E(\hat{N}_{hk}) \approx N_{hk}$ so that the $\hat{N}_{hk}$ are approximately unbiased (i.e., consistent) estimators of N_{hk}.

One problem with the raking technique is that the weighted sample marginals may not converge to the population marginals or may converge only very slowly. Although there does not seem to exist an easily verifiable set of necessary and sufficient conditions to determine whether convergence will occur (Lessler & Kalsbeek

1992:199), one rule of thumb suggested by Oh and Scheuren (1983:168) is not to use raking unless "the effective sample size" (i.e., in the case of non-response adjustment $n_{h\cdot}$ or $n_{\cdot k}$) is greater than 20 or 25 times the number of dimensions being constrained. For example, suppose a 4×4 table (i.e., 2 dimensions) is being raked. The suggestion seems to be not to use raking unless all the row and column totals are greater than $2 \times 20 = 40$ or $2 \times 25 = 50$. Oh and Scheuren (1983:168) also warn against imposing too many constraints on the sample (i.e., raking a large dimension or having a large number of adjustment cells *HK*), imposing constraints that are themselves contradictory or raking cells with very small expected sample sizes. Furthermore, if there are sizeable differences between the expected sample totals and the population totals, convergence may take a very long time (Oh & Scheuren 1978b:718).

5.6.1.3. Example of Raking Ratio Estimation

Suppose the sample frequencies over two cross-classified survey variables, each with three categories, are as follows:

k \ h	1	2	3	Total
1	10	20	15	45
2	20	40	30	90
3	15	35	25	75
Total	45	95	70	210

Note that all marginals are greater than $2 \times 20 = 40$. Suppose the known population marginals for these variables are:

k \ h	1	2	3	Total
1				150
2				650
3				200
Total	100	600	300	1000

The aim of the raking procedure is to adjust the sample counts n_{hk} so that estimates are obtained of the unknown population cell frequencies N_{hk}. In step 1 of the raking procedure, the row weighted cell values are obtained from (5.48), namely:

$$n_{1k}^{(1)} = \frac{150}{45} n_{1k} = 3{,}333 n_{1k} = a_1^{(1)} n_{1k}$$

$$n_{2k}^{(1)} = \frac{650}{90} n_{2k} = 7{,}222 n_{2k} = a_2^{(1)} n_{2k}$$

$$n_{3k}^{(1)} = \frac{200}{75} n_{3k} = 2{,}667 n_{3k} = a_3^{(1)} n_{3k}$$

for each column $k = 1, 2, 3$. The row-adjusted values are:

h / k	1	2	3	Total
1	33,333	66,667	50	$150 = N_{1.}$
2	144,444	288,889	216,667	$650 = N_{2.}$
3	40,000	93,333	66,667	$200 = N_{3.}$
Total	217,778	448,889	333,333	1000

Note that the row marginals have now been scaled to the known population marginals. In step 2, the column-adjusted values are obtained from (5.51):

$$\begin{aligned} n_{h1}^{(2)} &= \frac{100}{217{,}778} n_{h1}^{(1)} = 0{,}459 n_{h1}^{(1)} \\ &= a_h^{(1)} \times 0{,}459 n_{h1} \end{aligned}$$

$$\begin{aligned} n_{h2}^{(2)} &= \frac{600}{448{,}889} n_{h2}^{(1)} = 1{,}337 n_{h2}^{(1)} \\ &= a_h^{(1)} \times 1{,}337 n_{h2} \end{aligned}$$

$$\begin{aligned} n_{h3}^{(2)} &= \frac{300}{333{,}333} n_{h3}^{(1)} = 0{,}9 n_{h3}^{(1)} \\ &= a_h^{(1)} \times 0{,}9 n_{h3} \end{aligned}$$

for each row $h = 1, 2, 3$. The adjusted frequencies at the end of the first cycle of the raking procedure are:

k \ h	1	2	3	Total
1	15,298	89,124	44,996	149,418
2	66,298	386,233	194,994	647,525
3	18,362	124,802	60,008	203,172
Total	$99{,}958 \approx N_{.1}$	$600{,}159 \approx N_{.2}$	$299{,}998 \approx N_{.3}$	1000

Note that the column marginals have now been scaled to the known population marginals. (The population column marginals in this example are obtained only approximately due to rounding errors.)

In the second cycle, the procedure returns to step 1 applied to the $n_{hk}^{(2)}$ values and continues to iterate until the process converges. In this example, the values converge on the marginal totals at the end of cycle 2. The final estimates are:

k \ h	1	2	3	Total
1	15,360	89,466	45,174	150
2	66,562	387,687	195,751	650
3	18,078	122,847	59,075	200
Total	100	600	300	1000

5.6.2. Raking Adjustment for Non-response

Raking ratio estimation has also been used to adjust for non-response although it is a more sophisticated and less familiar technique than population or sample weighting adjustments. Raking ratio adjustment for non-response requires the formation of *adjustment cells* formed by the cross-classification of two or more (categorical) auxiliary variables in a multi-way table. The marginal (population) totals of the auxiliary variables are known from sources external to the survey. Raking is a useful non-response adjustment technique, because (Kalton 1983a:56):

1. It avoids part of the variance increases produced by the formation of a large number of weighting classes in population and sample weighting adjustments

2. Only the population *marginal* frequencies and not the *joint* frequencies of the auxiliary variables need to be known but the sample sizes within each adjustment cell must be known

5.6.2.1. Assumptions

In raking adjustment for non-response, the population is assumed to consist of L sub-populations formed by the cross-classification of, say, two auxiliary variables with H and K categories respectively ($L = HK$). The number of respondents in each sub-population is denoted as $n_{r_{hk}}$ where $h = 1, \ldots, H$ and $k = 1, \ldots, K$. Raking adjustment for non-response requires the assumptions that (1) population elements within each sub-population have independent and common response probabilities $\varphi_{hk} > 0$ (hence, respondents and non-respondents are similar within adjustment cells) and (2) the response probabilities are independent from one sub-population to another. These assumptions are the same as those required for sample and population weighting adjustments. Raking adjustment for non-response requires the *additional* assumption (3) that the response probabilities φ_{hk} are determined solely by the row or column an element falls in and do not depend also on a particular cell which implies that $\varphi_{hk} = \varphi_h\varphi_k$ (i.e., there is no interaction in the table to be raked) (Oh & Scheuren 1983:164).

Oh and Scheuren (1983:164) state that the requirement $n_{r_{hk}} > 0$ does not have to be satisfied in raking adjustment (contrary to population and sample weighting adjustments). However, the patterns of zeros among the $n_{r_{hk}}$ must be such that the algorithm converges, for example, the $n_{r_{hk}}$ cannot all be zero for a particular row or column. On the other hand, Little and Rubin (1987:60) believe that the raking ratio estimator is not defined when $n_{r_{hk}} = 0$ but $n_{hk} \neq 0$ (see section 5.6.2.2).

5.6.2.2. The Estimator of the Population Total and its Properties

Suppose the raking procedure is terminated (due to convergence or achievement of the required degree of closeness[10]) at the end of cycle c. The raking ratio estimator of the population total is (Oh & Scheuren 1983:165):

---

[10] Oh and Scheuren (1983:169) have found that the use of *convergent* raking estimates as opposed to limited raking estimates lead to a greater reduction in the bias of estimators but at the price of some increase in the variance.

$$\hat{Y}_q = \sum_{h=1}^{H}\sum_{k=1}^{K} \hat{N}_{hk}^{(c)} \bar{y}_{r_{hk}} \tag{5.56}$$

where $\hat{N}_{hk}^{(c)} = \omega_{hk}^{(c)} n_{hk}$ and $\omega_{hk}^{(c)}$ is obtained from (5.55). According to Little and Rubin (1987:60) some other estimator of the mean of a cell is required when $n_{r_{hk}} = 0$, but $n_{hk} \neq 0$.

The conditional bias of $\hat{Y}_q$ given $n_{r_{hk}}$ (and given n_{hk} as before) is (Oh & Scheuren 1983:165):

$$Bias(\hat{Y}_q | n_{r_{hk}}) = -\sum_{h=1}^{H}\sum_{k=1}^{K} (\bar{Y}_{hk} - \bar{Y})(N_{hk} - \hat{N}_{hk}^{(c)}) \tag{5.57}$$

The bias of the raking ratio estimator is, therefore, dependent on the magnitude of the differences in the population means among the cells and on the degree of closeness of the raked marginals to the population marginals. Thus, if $N_{hk} \approx \hat{N}_{hk}^{(c)}$ for all h and k, then $Bias(\hat{Y}_q | n_{r_{hk}}) \approx 0$. Since:

$$(\bar{Y}_{hk} - \bar{Y}) = \left(\bar{Y}_{h\cdot} - \bar{Y}\right) + \left(\bar{Y}_{\cdot k} - \bar{Y}\right) + \left(\bar{Y}_{hk} - \bar{Y}_{h\cdot} - \bar{Y}_{\cdot k} + \bar{Y}\right) \tag{5.58}$$

the conditional bias may also be expressed as (Oh & Scheuren 1983:165):

$$\begin{aligned} Bias(\hat{Y}_q | n_{r_{hk}}) &= \sum_{h=1}^{H}\left(\bar{Y}_{h\cdot} - \bar{Y}\right)\left(N_{h\cdot} - \hat{N}_{h\cdot}^{(c)}\right) + \sum_{k=1}^{K}\left(\bar{Y}_{\cdot k} - \bar{Y}\right)\left(N_{\cdot k} - \hat{N}_{\cdot k}^{(c)}\right) \\ &+ \sum_{h=1}^{H}\sum_{k=1}^{K} (\bar{Y}_{hk} - \bar{Y}_{h\cdot} - \bar{Y}_{\cdot k} + \bar{Y})(N_{hk} - \hat{N}_{hk}^{(c)}) \end{aligned} \tag{5.59}$$

Since $E\left(\hat{N}_{hk}^{(c)}\right) \approx N_{hk}$, and $E\left(\hat{N}_{h\cdot}^{(c)}\right) \approx N_{h\cdot}$ and $E\left(\hat{N}_{\cdot k}^{(c)}\right) \approx N_{\cdot k}$ the estimator $\hat{Y}_q$ is approximately unbiased for large samples.

The conditional variance of $\hat{Y}_q$ given $n_{r_{hk}}$ for sampling without replacement from a finite population may be written as (Oh & Scheuren 1983:166):

$$\begin{aligned} V(\hat{Y}_q | n_{r_{hk}}) &= \sum_{h=1}^{H}\sum_{k=1}^{K} (\hat{N}_{hk}^{(c)})^2 \left(1 - \frac{n_{r_{hk}}}{N_{hk}}\right) \frac{S_{hk}^{\ 2}}{n_{r_{hk}}} \\ &= \sum_{h=1}^{H}\sum_{k=1}^{K} \left[\left(\omega_{hk}^{(c)}\right)^2 n_{r_{hk}}\right] \left(1 - \frac{n_{r_{hk}}}{N_{hk}}\right) S_{hk}^{\ 2} \end{aligned} \tag{5.60}$$

A disadvantage of the raking estimator for non-response is that, because of its complexity, the approximate *MSE* can be extremely difficult to estimate (Oh & Scheuren 1983:168). Furthermore, Oh and Scheuren (1983:166) describe two situations where the population weighted estimator $\hat{Y}_{pw}$ may be preferred above the raking estimator $\hat{Y}_q$:

1. Suppose the variance S_{hk}^2 is large in a particular sub-population relative to the remaining population and N_{hk} is also large for this sub-population, then whenever $\hat{N}_{hk}^{(c)} > N_{hk}$ it is possible that $V(\hat{Y}_q \mid n_{r_{hk}}) > V(\hat{Y}_{pw} \mid n_{r_h})$.

2. If the value of $n_{r_{hk}}$ is non-zero whenever N_{hk} is non-zero, then $\hat{Y}_{pw}$ is conditionally unbiased under the model of data missing at random, while $\hat{Y}_q$ in general is not. In this case, even when $V(\hat{Y}_q \mid n_{r_{hk}}) < V(\hat{Y}_{pw} \mid n_{r_h})$ it is possible that $MSE(\hat{Y}_q \mid n_{r_{hk}}) > MSE(\hat{Y}_{pw} \mid n_{r_h})$.

Furthermore, $\hat{Y}_{pw}$, unlike $\hat{Y}_q$, does not depend for its unbiasedness on the assumption that $\varphi_{hk} = \varphi_h\varphi_k$ (Oh & Scheuren 1983:166).

Although primarily described here in terms of marginal "population weighting" adjustments, raking can also be applied with sample weighting adjustments: the marginal *sample* frequencies of the auxiliary variables are determined for the total sample of respondents and non-respondents combined, i.e., $n_{h\cdot}$ and $n_{\cdot k}$ and then the raking algorithm is applied as above, i.e., $N_{h\cdot}$ in Table 5.4 is replaced by $n_{h\cdot}$ and $n_{k\cdot}$ in Table 5.5 is replaced by $n_{r_{h\cdot}}$.

5.6.2.3. Choice of Adjustment Cells

The same kinds of bias-variance trade-offs exist in raking as in population or sample weighted estimation. The adjustment cells should be formed in such a way that the assumption of a uniform within-cell response probability is tenable. Furthermore, the auxiliary variables used in the cross-classification should be related to the (primary) survey variable(s) (Oh & Scheuren 1983:167).

5.6.2.4. Two Examples of Raking Ratio Adjustment for Non-response

In this section, raking adjustments for non-response are illustrated by continuing Examples 1 and 2 in section 5.5. Suppose that the population joint frequencies N_{hk} of the two auxiliary variables *race* and *urbanity* are not known but that the population marginals $N_{h.}$ and $N_{.k}$ are known. The two-way table of population frequencies are from Table 5.1:

Table 5.6 ***Unknown values of N_{hk} and known values of $N_{h.}$ and $N_{.k}$***

	Rural	Urban	Total
White	950	1700	2650
Black	2750	4600	7350
Total	3700	6300	10000

The respondent data from the sample yields the following frequencies (from Table 5.1):

Table 5.7 ***Values of $n_{r_{hk}}$***

	Rural	Urban	Total
White	36	135	171
Black	108	400	508
Total	144	535	679

In this example, the values $\hat{N}_{hk}$ converge at the end of two cycles. The final estimates are:

Table 5.8 ***Raking ratio estimates of N_{hk}***

	Rural	Urban	Total
White	974,864	1675,136	2650
Black	2725,136	4624,864	7350
Total	3700	6300	10000

5.6.2.4.1.Example 1

From Example 1 (Table 5.1) in section 5.5, the sample yields the following observed respondent means in the four adjustment cells:

Table 5.9 *Values of $\bar{y}_{r_h}$ for Example 1*

	Rural	Urban
White	R110	R150
Black	R20	R86

Using the raking ratio estimates of N_{hk} in Table 5.8, the raking ratio estimate of the population total is from (5.56):

$$\hat{Y}_q = \sum_{h=1}^{H}\sum_{k=1}^{K} \hat{N}_{hk}^{(c)} \bar{y}_{r_{hk}}$$
$$= R810746{,}46$$

Hence, the raking ratio estimate of the population mean is:

$$\hat{\bar{Y}}_q = \frac{1}{N}\hat{Y}_q = R81{,}07.$$

In Example 1 in section 5.5, the "true" population mean was given as $\bar{Y} = R81{,}45$, i.e., the population total is $Y = R814450$. Hence, the bias of the raking ratio estimate of the population mean is:

$$Bias(\hat{\bar{Y}}_q) = R81{,}07 - R81{,}45 = -R0{,}38$$

Raking has successfully reduced the non-response bias of the unadjusted estimate $\bar{y}_r$ ($Bias(\bar{y}_r) = R8{,}06$). It represents an improvement over the sample weighted estimate ($Bias(\hat{\bar{Y}}_{sw}) = -R1{,}75$), and a slight improvement over the population weighted estimate ($Bias(\hat{\bar{Y}}_{pw}) = -R0{,}44$).

5.6.2.4.2.Example 2

From Example 2 (Table 5.2) in section 5.5, the observed respondent means in the adjustment cells are:

Table 5.10 Values of $\bar{y}_{r_h}$ for Example 2

	Rural	Urban
White	R25 460	R35 880
Black	R7128	R11 160

The raking ratio estimator of the population total is from (5.56):

$$\hat{Y}_q = \sum_{h=1}^{H}\sum_{k=1}^{K} \hat{N}_{hk}^{(c)} \bar{y}_{r_{hk}}$$
$$= 1{,}55962 \times 10^8$$

Hence, the raking ratio estimator of the population mean is:

$$\hat{\bar{Y}}_q = \frac{1}{N}\hat{Y}_q = R15596{,}22$$

In Example 2 of section 5.5, the "true" population mean was given as $\bar{Y} = R18217$, i.e., the population total is $Y = 1{,}82167 \times 10^8$. Hence, the bias of the raking ratio estimate of the population mean is:

$$Bias(\hat{\bar{Y}}_q) = R15596 - R18217 = -R2621$$

In this example, raking adjustment has actually resulted in an increase in non-response bias relative to the unadjusted estimator ($Bias(\bar{y}_r) = -R2025$). As can be expected (since the population marginals are known), it is an improvement over the sample weighted estimator ($Bias(\hat{\bar{Y}}_{sw}) = -R3294$) and a *slight* improvement over the population weighted estimator ($Bias(\hat{\bar{Y}}_{pw}) = -R2605$).

The same remarks apply as in the case of population and sample weighting adjustments: raking ratio adjustment was not very successful in reducing non-response bias because of the choice of adjustment cells in this example. Adjustment cells should ideally be formed so that homogeneity within each adjustment cell is obtained.

5.7. OTHER WEIGHTING TECHNIQUES

5.7.1. Sub-sample Weighting for Non-response

Sub-sample weighting techniques are among the earliest weighting adjustments for non-response (Oh & Scheuren 1983:171). In the previous sections, respondents in the sample or in a particular weighting class or adjustment cell had fixed weights ω conditional on n and n_r. With sub-sample weighting, an additional level of sampling is introduced so that the weights of respondents in the sample or in specified classes or cells may vary. These techniques are related to some of the hot-deck imputation techniques discussed in Chapter 6. Two sub-sampling schemes that can be used to adjust for non-response are duplication weighting and integerisation weighting.

5.7.1.1. Duplication Weighting for Non-response

Duplication weighting involves the selection of a simple random sample of size n_{nr} from the n_r respondents (within the sample or within a weighting class) and using these additional elements in place of the n_{nr} non-respondents. The original estimation procedure is then carried out as if there had been complete response. For example, in the case of simple random sampling assuming the uniform global response mechanism, the duplication estimator can be written as (Oh & Scheuren 1983:172):

$$\hat{Y}_{dw} = \frac{N}{n}\left[\sum_{i=1}^{n_r} y_{r_i} + \sum_{i=1}^{n-n_r} y_{r_i}^{*}\right] \tag{5.61}$$

where the $y_{r_i}^{*}$ are a simple random sub-sample of size $n-n_r$ selected from the n_r responses y_{r_i}.

Suppose a simple random sample of $n = 30$ is selected from a population of size $N = 6000$ and $n_r = 20$ elements respond. Then $n_{nr} = n - n_r = 10$ respondents would have the weight $\omega_i = \dfrac{6000}{30}$ while the other 10 would have the weight twice as large. There are of course ${}_{20}C_{10}$ possible ways in which these weights could be assigned to the respondents (Oh & Scheuren 1983:172). (If $n_{nr} > n_r$ sampling must be done with replacement.)

It can be shown that the conditional variance of the estimator $\hat{Y}_w = N\bar{y}_r$ is always less than or equal to that of the duplication estimator $\hat{Y}_{dw}$ (Oh & Scheuren 1983:173). A question which arises immediately is why use an estimator which always has a larger variance than its "natural competitor"? The reason is that a self-weighting sample is often required. The duplication estimator will retain the self-weighting properties of a simple random sample or of any other self-weighting design no matter how many weighting classes are used as long as there is at least one respondent in each class for which there are non-respondents. (Oh & Scheuren 1983:173.)

5.7.1.2. Integer Weighting for Non-response

Integer weighting, like duplication weighting, attempts to simplify the adjusted weights of respondents by an additional randomisation step. The non-response adjusted weights (ω_i) are corrected to integers by a two-step process (Oh & Scheuren 1983:172):

1. Let $\omega_i = I_i + \theta_i$ where I_i is the largest integer less than or equal to ω_i and $0 \leq \theta_i < 1$ is the difference between ω_i and I_i.
2. Replace each θ_i by an indicator random variable B_i such that $P(B_i = 1) = \theta_i$.

This means that the final weight for the i-th respondent will be either I_i or $I_i + 1$. For example, consider the constant weighted estimator:

$$\hat{Y}_w = \frac{N}{n_r}\sum_{i=1}^{n_r} y_{r_i} = \sum_{i=1}^{n_r} \omega_i y_{r_i}$$

The integerised estimator adjusts the weights ω_i to obtain the estimator:

$$\hat{Y}_{Iw} = \sum_{i=1}^{n_r} (I_i + B_i) y_{r_i} \tag{5.62}$$

The integerised estimator has the same expected value as the original estimator; hence, if the original estimator was unbiased, $\hat{Y}_{Iw}$ is also unbiased. The integerised estimator, like the duplication estimator, tends to increase the variance although this effect may be quite small (Oh & Scheuren 1983:174). One advantage that integer weights provide is that they "assure consistency in weighted counts within (and among) multiple cross-tabulations and thus allow one to avoid the otherwise ubiquitous note: *details may not add to totals due to rounding*" (Oh & Scheuren 1983:174).

5.7.2. The Politz-Simmons Procedure

An early attempt at using estimated response probabilities to adjust for non-response due to Politz and Simmons (1949) has become a well-known but not widely used adjustment procedure. In this technique, *response probability* is equated to *availability* during interviewing hours. Personal interviews are scheduled at randomly determined times on the six evenings Monday to Saturday. During the interview, respondents are asked on how many of the previous five evenings they were at home at about the same time. If the answer is h nights, the quantity $\frac{h+1}{6}$ is taken as an estimate of the response probability φ_h. The sample results are then post-stratified into the six "strata" $h = 0, 1, \ldots, 5$. The Politz-Simmons estimator of the population mean is (Cochran 1977:374):

$$\hat{\bar{Y}}_{PS} = \frac{\sum_{h=0}^{5} \frac{6n_h \bar{y}_h}{h+1}}{\sum_{h=0}^{5} \frac{6n_h}{h+1}} \tag{5.63}$$

where n_h is the number of interviews obtained and $\bar{y}_h$ is the sample mean in the h-th "stratum". This estimator is less biased than the sample mean from the first call, but its variance is greater because an unweighted mean is replaced by a weighted mean with estimated weights (Cochran 1977:374).

An advantage of the Politz-Simmons technique is the saving in time since call-backs are eliminated by using weighted first-call data only. On the other hand, the technique has a number of disadvantages:

1. A strong assumption must be made that the respondents supply correct values of h
2. It is difficult in practice to enforce the randomisation of the time of the actual call
3. The procedure fails to account for those who are absent from home during all six evenings since none of them are interviewed

4. The Politz-Simmons procedure is designed to deal with not-at-homes, not refusals

According to Kish (1965:560), the Politz-Simmons technique has serious problems of validity and practicality. Nevertheless, Kish (1965:560) states that if the survey situation permits only a single call, the Politz-Simmons estimate, despite its bias and increased variance, may be preferable to accepting the bias of unweighted first calls.

A few other methods have been proposed to identify reasonable estimates of the response probabilities φ_i, but have not been widely adopted as adjustment procedures (Kalton & Kasprzyk 1986:5). (See for example Chapman (1976), Drew and Fuller (1980, 1981) and Thomsen and Siring (1983)).

5.7.3. Linear Regression Estimators

According to Bethlehem and Keller (1987:141), there are two major problems with the use of the population weighted estimator, namely empty weighting classes and lack of adequate population information. They suggest a general weighting method which avoids the mentioned problems but take advantage of all the available auxiliary information. The weights are obtained from a linear model which relates the survey variable to a set of auxiliary variables. The theory of the *non-response adjusted generalised linear regression estimator* will not be discussed in detail here but further information on the estimator may be obtained in Bethlehem and Kersten (1985), Bethlehem (1988), Bethlehem and Keller (1987) and Särndal and Swensson (1985, 1987).

5.7.3.1. Assumptions

For simplicity, only the case of the uniform global response mechanism will be considered in this section. The use of the regression estimator in the case of the uniform response mechanism within sub-populations is discussed by Särndal and Swensson (1985, 1987). Furthermore, attention in this section is no longer restricted to the case of a simple random sample, but in general to a probability sample selected without replacement from the finite population and with inclusion probabilities π_i for $i = 1, \ldots, N$.

5.7.3.2. The Generalised Regression Estimator in the Case of Full Response

Suppose as before that there is a single survey variable y and suppose the aim of the survey is to estimate the population mean $\bar{Y}$. Recall from Chapter 1 that the Horvitz-Thompson estimator:

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{j=1}^{N}\frac{y_j}{\pi_j} \tag{5.64}$$

is an unbiased estimator of the population mean in the case of full response.

Suppose now there are p auxiliary variables, $x_1, \ldots, x_p$ correlated with y, whose values are known for all population elements. Each element in the population can thus be associated with a $1 \times p$ vector of auxiliary variable values $\underline{\mathbf{x}}'_j = \left(x_{j1}, x_{j2}, \ldots, x_{jp}\right)$ for $j = 1, \ldots, N$. The notation $\underline{\mathbf{x}}$ or $\underline{\mathbf{X}}$ will be used to denote a column vector and the notation $\underline{\mathbf{x}}'$ or $\underline{\mathbf{X}}'$ will be used to denote a row vector. The $N \times p$ matrix of all values of the auxiliary variables is denoted by $\mathbf{X}$. The population mean vector for the p auxiliary variables is denoted by $\bar{\underline{\mathbf{X}}}$.

According to Bethlehem and Keller (1987:142) the precision of the Horvitz-Thompson estimator can be improved by using the regression coefficients obtained from a linear regression of y on the auxiliary variables. Let $\underline{\beta}' = \left(\beta_1, \beta_2, \ldots, \beta_p\right)$ denote the vector of ordinary least squares regression coefficients for the best fit of y on $\underline{\mathbf{x}}$, where:

$$\underline{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{y}} = \left[\sum_{j=1}^{N}\underline{\mathbf{x}}_j\underline{\mathbf{x}}'_j\right]^{-1}\left[\sum_{j=1}^{N}\underline{\mathbf{x}}_j y_j\right] \tag{5.65}$$

In the case of full response, $\underline{\beta}$ can be estimated by the approximately unbiased (consistent) estimator (for large samples) (Bethlehem 1988:255):

$$\hat{\underline{\beta}} = \left[\sum_{i=1}^{n}\frac{\underline{\mathbf{x}}_i\underline{\mathbf{x}}'_i}{\pi_i}\right]^{-1}\left[\sum_{i=1}^{n}\frac{\underline{\mathbf{x}}_i y_i}{\pi_i}\right] \tag{5.66}$$

The generalised regression estimator of the population mean in the case of full response is defined as:

$$\hat{\bar{Y}}_{GR} = \hat{\bar{Y}}_{HT} + (\underline{\bar{\mathbf{X}}} - \underline{\hat{\bar{\mathbf{X}}}}_{HT})'\hat{\underline{\beta}} \tag{5.67}$$

where $\hat{\underline{\beta}}$ is obtained from (5.66) and $\underline{\hat{\bar{\mathbf{X}}}}_{HT}$ is the vector of Horvitz-Thompson estimators of the p auxiliary variables defined in the same way as (5.64) (Bethlehem 1988:255). The estimator $\hat{\bar{Y}}_{GR}$ is approximately unbiased (consistent) for large samples.

5.7.3.3.The Non-response Adjusted Linear Regression Estimator

In the presence of non-response and assuming the uniform global response mechanism, the Horvitz-Thompson estimators $\hat{\bar{Y}}_{HT}$ and $\underline{\hat{\bar{\mathbf{X}}}}_{HT}$ must be replaced by their non-response adjusted equivalents as defined in section 5.2. Furthermore, estimation of $\underline{\beta}$ will have to be based on the respondent data only, namely (Bethlehem 1988:256):

$$\hat{\underline{\beta}}^* = \left[\sum_{i=1}^{n_r} \frac{\underline{\mathbf{x}}_i \underline{\mathbf{x}}'_i}{\pi_i}\right]^{-1} \left[\sum_{i=1}^{n_r} \frac{\underline{\mathbf{x}}_i y_i}{\pi_i}\right] \tag{5.68}$$

The modified generalised regression estimator of the population mean is now defined as (Bethlehem 1988:256):

$$\hat{\bar{Y}}^*_{GR} = \hat{\bar{Y}}_{\pi^*} + (\underline{\bar{\mathbf{X}}} - \underline{\hat{\bar{\mathbf{X}}}}_{\pi^*})'\hat{\underline{\beta}}^* \tag{5.69}$$

where from (5.2):

$$\hat{\bar{Y}}_{\pi^*} = \frac{\sum_{i=1}^{n_r} \frac{y_{r_i}}{\pi_i \varphi_i}}{\sum_{i=1}^{n_r} \frac{1}{\pi_i \varphi_i}} \tag{5.70}$$

and $\underline{\hat{\bar{\mathbf{X}}}}_{\pi^*}$ is the analogue of (5.70)[11] (Bethlehem 1988:254).

In the case of a simple random sample the estimator of $\underline{\beta}$ is (Bethlehem & Kersten 1985:296):

$$\hat{\underline{\beta}}^* = \left[\sum_{i=1}^{n_r} \underline{\mathbf{x}}_i \underline{\mathbf{x}}'_i\right]^{-1} \left[\sum_{i=1}^{n_r} \underline{\mathbf{x}}_i y_i\right] \tag{5.71}$$

[11] Note that in the case of full response, estimator (5.70) does not reduce to the original Horvitz-Thompson estimator. Instead, the population size N is replaced by its estimator based on the sample.

The modified generalised regression estimator of the population mean is now defined as:

$$\hat{\bar{Y}}^{*}_{GR} = \hat{\bar{Y}}_{w} + (\underline{\bar{\mathbf{X}}} - \hat{\underline{\bar{\mathbf{X}}}}_{w})'\hat{\underline{\beta}}^{*} \tag{5.72}$$

where, from section 5.2.3.2:

$$\hat{\bar{Y}}_{w} = \frac{\frac{n}{n_r}\sum y_{r_i}}{\frac{n}{n_r}n_r} = \bar{y}_r \tag{5.73}$$

and $\hat{\underline{\bar{\mathbf{X}}}}_{w} = \underline{\bar{\mathbf{x}}}_r$ is the analogue of (5.73).

It can be shown that the use of the estimator implies that the weight assigned to the i-th sample element is given by (Bethlehem & Kersten 1985:296):

$$\omega_i = \underline{\mathbf{x}}'_i\left[\sum_{i=1}^{n_r}\underline{\mathbf{x}}_i\underline{\mathbf{x}}'_i\right]\underline{\bar{\mathbf{X}}} \tag{5.74}$$

5.7.3.4. Properties of the Non-response Adjusted Linear Regression Estimator

The expected value of $\hat{\bar{Y}}_{\pi^*}$ can be approximated by:

$$E(\hat{\bar{Y}}_{\pi^*}) \approx \bar{Y}^{*} = \frac{1}{N}\sum_{j=1}^{N}\frac{\varphi_j y_j}{\bar{\varphi}} \tag{5.75}$$

where $\bar{\varphi} = \frac{1}{N}\sum_{j=1}^{N}\varphi_j$ is the mean response probability in the population. Similarly, the expected value of $\hat{\underline{\bar{\mathbf{X}}}}_{\pi^*}$ can be approximated by:

$$E(\hat{\underline{\bar{\mathbf{X}}}}_{\pi^*}) \approx \underline{\bar{\mathbf{X}}}^{*} = \frac{1}{N}\sum_{j=1}^{N}\frac{\varphi_j \underline{\mathbf{x}}_j}{\bar{\varphi}} \tag{5.76}$$

so that the approximate bias of the modified regression estimator is equal to:

$$Bias(\hat{\bar{Y}}^{*}_{GR}) \approx \left(\bar{Y}^{*} - \bar{Y}\right) - \left(\underline{\bar{\mathbf{X}}}^{*} - \underline{\bar{\mathbf{X}}}\right)\underline{\beta}^{*} \tag{5.77}$$

where:

$$\underline{\beta}^* = \left(\sum_{j=1}^{N} \varphi_j \underline{\mathbf{x}}_j \underline{\mathbf{x}}'_j\right)^{-1} \left(\sum_{j=1}^{N} \varphi_j \underline{\mathbf{x}}_j y_j\right) \tag{5.78}$$

The bias vanishes if $\underline{\beta}^*$ is equal to $\underline{\beta}$. Thus, if non-response does not affect the regression coefficients, the resulting regression estimator will be unbiased. $\underline{\beta}^*$ and $\underline{\beta}$ will be equal if:

1. there is no correlation between the residuals of the regression model and the response behaviour and
2. the residuals are small, i.e., if there is a good fit (Bethlehem 1988:256).

The vector of residuals $\underline{\varepsilon}' = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$ is defined as (Bethlehem 1988:255):

$$\underline{\varepsilon} = \underline{y} - \mathbf{X}\underline{\hat{\beta}}^* \tag{5.79}$$

5.8. CONCLUSION

According to Oh and Scheuren (1983:144), the response models employed in adjusting for non-response, no matter how cleverly structured, virtually never hold in practice. As Kalton (1983b:182) puts it:

> *... sampling practitioners do not believe that the nonresponse models on which their adjustments are based hold exactly: they simply hope that they are improvements on the model of data missing at random over the total population implicitly assumed in estimating means with the "do nothing" procedure. In consequence, they hope that the nonresponse biases will be reduced, but they do not truly expect them to be entirely corrected.*

The probability response models employed in weighting adjustments for non-response are non-robust:

1. because they are unsuccessful in dealing with zero response probabilities (for example for hard-core non-respondents), and
2. because of their inability to model accurately the statistical dependence of the response probability on the auxiliary variables used to define the weighting or adjustment classes.

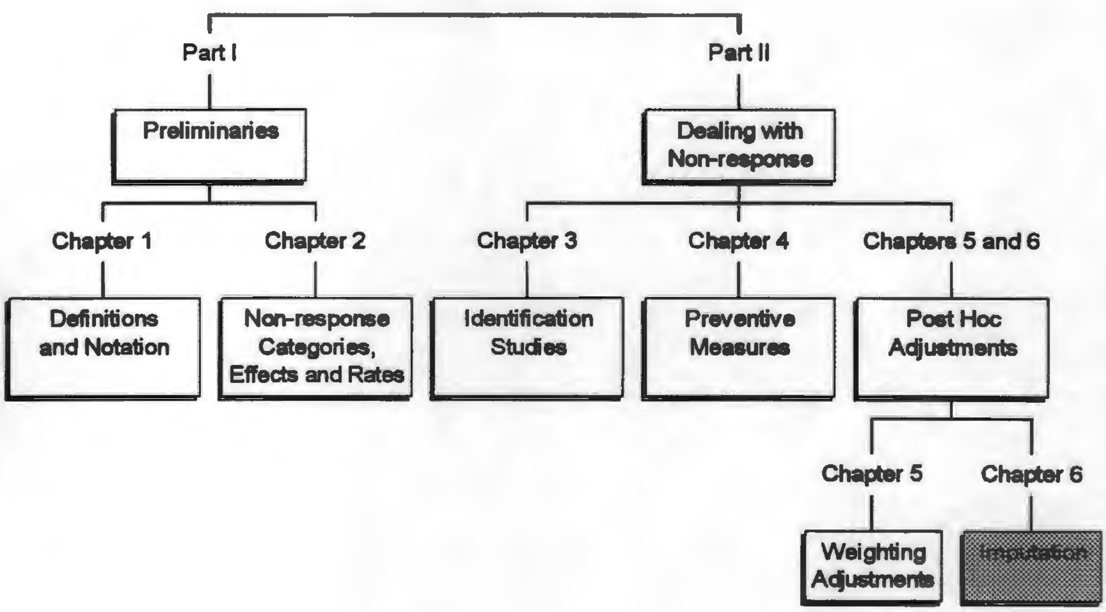
Oh and Scheuren (1983:144) remark that non-response always brings with it an increase in the *MSE*. Even if one models the response mechanism properly, there is an additional component of the variance due directly to the non-response (because of the reduced sample size). There can also be a further variance increase, depending on exactly how the adjustment is carried out.

According to Bailey (1983:294), an all-purpose non-response data adjustment technique is highly inconceivable: there is no single optimal non-response weighting adjustment. In reducing the bias for some variables, it is possible, even if the response mechanism is modelled correctly, to increase the bias of others. Oh and Scheuren (1983:158) suggest that practitioners employ more than one set of non-response weighting adjustments depending on the resources available, the amount of non-response, the degree of uncertainty about the response mechanism and the extent to which different purposes are to be served by the same data set.

CHAPTER 6

COMPENSATING FOR ITEM NON-RESPONSE:

IMPUTATION



CHAPTER OUTLINE

CHAPTER 6

COMPENSATING FOR ITEM NON-RESPONSE:

IMPUTATION

Reality does not consist of the data at the end of the chapter of some textbook (like the iris data) and normal distributions; it consists of 20 000 long forms filled out by 20 000 businessmen with other things on their minds, or several million census returns filled out by individuals who want to get back to the newspaper or the TV. These people want to be cooperative; but if the information requested is not handy or has been forgotten, they pass over the question or make up a response, and they also make mistakes. The survey people have to extract as much sense as possible from the results, and they try to do a respectable and ethical job.

Innis G. Sande, Imputation in Surveys: Coping with Reality

6.1. INTRODUCTION

In addition to unit non-response, most surveys are also subject to item non-response, i.e., responses for one or more (but not all) survey items are missing. A response to an item in a questionnaire may be "truly missing", i.e., no response was recorded for the item or it may be "artificially missing", i.e., a response was recorded but it was eliminated during the editing process because it did not satisfy natural or reasonable constraints or it was inconsistent with other responses (Sande 1982:145). There are at least four approaches to dealing with item non-response, namely:

1. Contact the respondent again and clear up the problem.
2. Discard the records with missing values in *analyses* but tabulate the number of missing values separately as a non-response category.
3. Calculate a separate set of weights (as described in Chapter 5) for each item with missing data, i.e., use weighting adjustments to compensate for *both* unit and item non-response.
4. Create a "clean" (i.e., a completely filled) rectangular data matrix before calculating estimates for each item, i.e., weighting adjustments for unit non-response are based on the clean data set. Such a "clean data matrix approach" involves *imputation* which can be defined as the process of assigning values $\hat{y}_i$ for each missing value y_i - exactly how the imputation value is calculated is not important to the definition (Ford 1983:186).

The fourth approach, i.e., imputation, is the preferred procedure for dealing with item non-response. Although the first approach, i.e., re-contacting the respondent to obtain complete or "correct" responses is the ideal solution, it is often impossible, impractical or too expensive and it does not solve the problem of refusals. Furthermore, problems with data are sometimes discovered when the survey processing is well advanced and it is too late to "stop the presses" while the respondent is consulted (Sande 1982:145).

The second approach, i.e., discarding the records with missing values, has an advantage in that secondary users of the data will be allowed to select their own methods of dealing with the missing data. However, this is not the ideal solution, for various reasons:

a) It fails to make use of the partial information about the missing values that may be available in other responses.

b) When forming cross-tabulations, the deletion of records with missing values is bound to lead to inconsistencies in marginal distributions between different tabulations involving the same variables. (Kalton 1983a:65)

c) It is better that the *data collector* instead of the *data analyst* or other users of the data make the non-response adjustments. The data collector usually has first-hand knowledge of the data which can be used to his/her advantage in adjusting for missing data. Furthermore, a clean data set allows all future analyses to have a common starting point without every analyst imputing his/her own values and getting results that are inconsistent with one another and the original (Ford 1983:189).

d) Many standard types of statistical analyses, such as regression analysis, require a clean data set. Some computer programs require a clean data set before any analyses can be performed and in others, each record with a missing value for an item is automatically excluded from analyses. This results in the loss of a large amount of data if many records are affected by item non-response. (Lehtonen & Pahkinen 1995:126.)

There are a number of reasons why imputation is preferred to the item-by-item adjustment approach:

a) The computation of separate weighting adjustments for each missing data item is usually a complex and time-consuming procedure. On the other hand, imputation makes analyses easier and the results simpler to present (Kalton 1983a:65).

b) When it comes to multivariate analyses involving more than one variable with missing data, the analyst faces the problem of which item-level weights to use (Lessler & Kalsbeek 1992:211).

c) Separate weighting adjustments may give rise to inconsistencies between estimates and tables: although a set of estimates is produced, no complete data matrix exists which satisfies all the edit constraints and corresponds to this set of estimates (Sande 1982:145).

Various imputation methods are discussed in this chapter. The technique of *deductive imputation*, discussed in section 6.2.2, involves using responses to related data items to logically determine the missing responses. *Overall mean* or *class mean imputation*, discussed in sections 6.2.3 and 6.2.4, involves replacing missing data values with the average value of respondent data in the sample or in specially defined sub-groups of the sample. The *cold-deck imputation* procedure, discussed in section 6.2.5, replaces missing values in the current survey with values from some previous survey. In contrast, *hot-deck imputation* procedures replace missing data values with responses in the current survey. Various types of hot-deck procedures are discussed in section 6.2.6, amongst others, *randomised* and *sequential* hot-deck procedures as well as *distance function matching*. In *regression imputation*, discussed in section 6.2.7, missing responses are replaced by a value predicted from a regression model. Instead of producing a single imputed value, it has been suggested that several complete data matrices be produced and a combination of imputed values from each data set be used to produce a final estimate of population values. The technique of *multiple imputations* is discussed in section 6.2.8.

6.2. IMPUTATION METHODS

Because most imputation methods originated in survey practice on a "common-sense" basis with little theory to direct their development, few definitions or theoretical results about these procedures are available (Ford 1983:187). According to Ford (1983:185), "widespread practice in the absence of well-developed theory clouds the subject with ambiguities and inconsistencies", and according to Bailar and Bailar (1983:301), "despite the increasing use and importance of imputation methods, little has been done to evaluate critically and to compare imputation techniques".

Against this background, a description of some of the main imputation methods will be given - but not of all the many variants of these methods which are used in practice. The bias and variance of estimators (of the population *mean* only) will be given for some, but not all of the imputation techniques discussed.

6.2.1. Notation and Assumptions

In this chapter, the problem of item non-response is separated from the problem of unit non-response by making the assumption that there is no unit non-response. Additional terminology introduced in this chapter, includes the terms *donor* and *recipient*. Many imputation methods assign the value from a record with a response on the item (the donor record) to a record with a missing value on the item (the recipient).

In practice, survey records often contain many missing values but for simplicity, imputation will be discussed here in terms of a *single item* with missing responses. Such a univariate approach to imputation presents some problems when multivariate analyses are to be performed (Lessler & Kalsbeek 1992:212). For example, if several missing values in a single record are imputed separately from several donors, the covariance structure is attenuated. It is preferable therefore, that these values are imputed jointly, using the same respondent as the donor, so that the covariance structure is retained (Kalton & Kasprzyk 1986:11). (See section 6.2.6.2.)

The notation used and the assumptions made in this chapter are similar to those in Chapter 5. Specifically, no further effort will be expended on a discussion of notation and model assumptions, except for the following two comments:

1. As before, inferences are made conditional on the values of y, n and n_r

2. Most of the imputation methods discussed are based on the same underlying model for the response mechanism, namely, a uniform response mechanism where the data are assumed to be missing at random, usually within sub-groups of the sample. Thus, the assumption is made that the non-respondents in the sample (or in specified sub-groups of the sample) follow the same distribution as the respondents.

6.2.2. Deductive Imputation

This (ideal) method of dealing with item non-response can be applied in situations, rare in practice, where the correct value for a missing response can be deducted logically from other information on the respondent (Kalton 1983a:68). Thus, for example, if a respondent's gender is missing but the respondent has a female name and/or the respondent's relationship to the head of the household is "wife", the gender of the respondent may be deduced (with *reasonable* certainty at least!) to be female. A respondent's race may also be inferred from, for example, the spelling of his/her surname or (maybe more so in the past) from his/her residential area.

Deductive imputation is often used as part of the editing process prior to other forms of imputation (Lessler & Kalsbeek 1992:224).

6.2.3. Overall Mean Imputation

This simple imputation method assigns the respondent item mean $\bar{y}_r$ as a substitute for each missing value of the item[1]. The rational for choosing $\bar{y}_r$ to impute for missing values is that the overall mean of the values in the clean data set:

$$y_1,\ y_2,\ \ldots,\ y_{n_r},\ \bar{y}_r,\ \bar{y}_r,\ \ldots,\ \bar{y}_r$$

then equals the mean of the true response values $\bar{y}_r$ (Lanke 1983:106). If, for simplicity, the first $n_r < n$ elements are labelled as respondents, the overall mean imputation estimator of the population mean assuming simple random sampling without replacement, can be written as:

[1] Other estimators of the centre of the distribution, such as the median or the mode in the case of categorical or nominal data may also be considered (Madow, Nisselson & Olkin 1983:87).

$$\begin{aligned}\hat{\bar{Y}}_M &= \frac{1}{n}\left[\sum_{i=1}^{n_r} y_{ri} + \sum_{i=1}^{n_{nr}} \bar{y}_r\right] \\ &= \frac{1}{n}\left[n_r\bar{y}_r + n_{nr}\bar{y}_r\right] = \bar{y}_r\end{aligned} \tag{6.1}$$

The overall mean imputation estimator is therefore equal to the constant weighting adjustment estimator $\hat{\bar{Y}}_w$ in (5.12). The estimator $\hat{\bar{Y}}_M$ will be unbiased under the model of data missing at random across the entire sample.

The variance of the estimator $\hat{\bar{Y}}_M$ is (Lanke 1983:106):

$$V(\hat{\bar{Y}}_M \mid n_r) = V(\bar{y}_r \mid n_r) = \left(1 - \frac{n_r}{N}\right)\frac{S^2}{n_r} \tag{6.2}$$

with an unbiased estimator of the variance under the model:

$$v(\hat{\bar{Y}}_M \mid n_r) = v(\bar{y}_r \mid n_r) = \left(1 - \frac{n_r}{N}\right)\frac{s_r^{\,2}}{n_r} \tag{6.3}$$

A *first* disadvantage of the overall mean imputation method is that it distorts the empirical distribution of the item: the concentration of all the imputed values at the mean creates a spike in the distribution (Kalton & Kish 1981:146). As a result, when population values other than means are being estimated, this method is likely to produce worse estimates than those derived from no imputation at all (Cox & Cohen 1985:225). For example, if the distribution of income is being estimated, mean value imputation will result in over-estimating the proportion of the population falling into the middle of the distribution and under-estimating the proportion with high and low incomes. This unsatisfactory heaping can be made less severe by imputing values based on the mean of a *sub-group* to which a record belongs (see section 6.2.4) or by choosing a donor value at random to replace a missing value (see section 6.2.6.1).

Secondly, as is intuitively clear, to replace all the missing values for a given item by the respondent mean for that item will give a set of n values with less variability than a sample of equal size consisting entirely of actually observed values (Särndal *et al.* 1992:592). The result is that over-optimistic results of the precision of estimators will emerge. Furthermore, standard computer programs are likely to produce the sample variance in the *complete* data set (with imputed values):

$$\hat{s}^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n_r} y_{r_i}^{\,2} + \sum_{i=1}^{n_{nr}} \bar{y}_r^{\,2} - n\bar{y}^2\right]$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n_r} y_{r_i}^{\,2} - n_r \bar{y}_r^{\,2}\right] \tag{6.4}$$
$$= \frac{n_r - 1}{n-1} s_r^{\,2}$$

(instead of $s_r^{\,2}$) as an estimator of S^2 and to estimate the variance of $\hat{\bar{Y}}_M$ as:

$$\hat{v}(\hat{\bar{Y}}_M \mid n_r) = \left(1 - \frac{n}{N}\right)\frac{\hat{s}^2}{n} \tag{6.5}$$

The computer program thus makes two mistakes: (1) it under-estimates S^2, since $\hat{s}^2 < s_r^{\,2}$ and $n_r < n$ (see (6.4)) and (2) it over-estimates the number of actual responses obtained. (Lanke 1983:106.)

6.2.4. Class Mean Imputation

As was shown in Chapter 5, the model of data missing at random across the entire sample can be improved by assuming a random distribution of missing values in specific sub-groups of the sample. Instead of imputing the overall respondent mean, most mean-imputation methods start with the division of the sample into H mutually exclusive, exhaustive and homogeneous sub-groups called *imputation classes*[2]. Missing values within an imputation class are then replaced by the respondent mean $\bar{y}_{r_h}$ in that class.

Imputation classes are analogous to weighting classes formed for weighting adjustments (see Chapter 5)[3]. They are formed from the cross-classification of "auxiliary" variables that are presumed to be highly correlated with the missing data item and are available for both respondents and non-respondents. For simplicity, the variables used to form imputation classes will be called auxiliary variables, although they may also be items (survey variables) with complete responses in the present survey.

---

[2] As described in Chapter 3, algorithms developed to produce homogeneous groupings of categorical variables such as CHAID, can be used to define imputation classes.

[3] Although imputation classes are similar to weighting classes for unit non-response, the range of choice for forming imputation classes is far greater than that for weighting classes because many more data are available for sample elements with item non-response.

The class mean imputation estimator of the population mean under simple random sampling without replacement can be written as:

$$\hat{\bar{Y}}_{CM} = \sum_{h=1}^{H} \frac{n_h}{n} \hat{\bar{y}}_h \tag{6.6}$$

where $\hat{\bar{y}}_h$ is the mean of class h of the clean data set, namely:

$$\hat{\bar{y}}_h = \frac{1}{n_h}\left[n_{r_h}\bar{y}_{r_h} + n_{nr_h}\bar{y}_{r_h}\right] = \bar{y}_{r_h} \tag{6.7}$$

Hence, the class mean imputation estimator of the population mean is equal to the sample weighted estimator $\hat{\bar{Y}}_{rw}$ in (5.19). This estimator will be unbiased under the model of data missing at random within each imputation class.

The variance of the estimator $\hat{\bar{Y}}_{CM}$ is:

$$V(\hat{\bar{Y}}_{CM} \mid n_r) = \sum_{h=1}^{H} w_h^{\,2}\left(1 - \frac{n_{r_h}}{N_h}\right)\frac{S_h^{\,2}}{n_{r_h}} \tag{6.8}$$

No variance estimator is given in the sources consulted but a variance estimator can be constructed by using $s_{r_h}^{\,2}$ as an estimator of $S_h^{\,2}$ and estimating N_h, if it is unknown, by $\frac{n_h}{n}N$.

Class mean imputation has the same disadvantages as the method of overall mean imputation, namely, the "natural" distribution of the sampled values is distorted (although the distortion is less severe than with overall mean imputation) and standard variance formulas will under-estimate the true variance (Särndal *et al.* 1992:592). More authentic variability in the imputed values is sought by methods which impute actual respondent values to non-respondents (see section 6.2.6).

A further disadvantage of the method of mean-value imputation (both overall mean and class mean imputation) is that is requires two passes through the data file: first in order to compute $\bar{y}_r$ or $\bar{y}_{r_h}$ then once more in order to perform the imputations. Ford (1983:195) explains why it is advantageous to employ an imputation method that requires a single pass through the data file:

> *Generally, computer costs and programming complexity are minimized by the "one record at a time principle" - the computer reads a data record, processes that record, reads a second record into the computer space occupied by the first record, processes the second record, etc. Thus, the computer "forgets" the first record before it processes the second. This principle is usually followed during the computer summarization of means, totals, standard errors, and other common statistics from even the most complex sample designs ... The effect of this "one record at a time principle" is still a major factor in the summarization of large-scale surveys.*

Ford (1983:196) does admit, however, that advances in data processing and innovative programming are overcoming this constraint.

A further problem with class mean imputation (and any other imputation methods which make use of imputation classes) involves the choice of auxiliary variables employed in defining imputation classes. If the auxiliary variables are continuous, they first need to be categorised - which necessarily involves some loss of information. Although the loss of information need not be severe provided a reasonable number of well-chosen categories is used, the method of class mean imputation generally is not ideal for continuous auxiliary variables (Little & Rubin 1987:65). The categorisation of continuous auxiliary variables can be avoided by the use of alternative imputation methods, such as methods based on minimising a distance function between donor and recipient in terms of the auxiliary variables (see section 6.2.6.5) or methods using regression for determining the imputed values (see section 6.2.7).

6.2.5. Cold-deck Imputation

The cold-deck method was an early form of computerised imputation which has now been superseded by what is known as "hot-deck" imputation (Kalton 1983a:69). The cold-deck procedure begins by forming imputation classes and assigning values to each class in advance of the survey. These values are usually responses obtained from a previous survey taken from essentially the same population or, in the case of a periodic survey, responses obtained on the previous round of the survey. There must be at least one "response" in each class. Once the cold-deck has been established and stored in the memory of the computer, the current survey records are considered in turn. Each record is associated with a specific imputation class. If a record has a missing response for the item, it is assigned a value from its imputation class in the cold-deck. If more than one value is stored in the imputation class, the imputed value may be selected at random or systematically (Chapman 1976:245).

An obvious disadvantage of the cold-deck method is that the historical values inserted into the cold-deck may not be comparable with the present survey responses. The solution to this problem is to employ a method based on current responses rather than past values for the imputation process. This leads to hot-deck imputation methods.

6.2.6. Hot-deck Imputation

The term "hot-deck" is used to describe a family of imputation methods which are widely used in current survey practice. Although a universally acceptable definition for this approach to imputation has not yet been suggested, a hot-deck method is generally one in which each missing value is replaced by a donor value from a similar respondent *in the same survey* (Lessler & Kalsbeek 1992:213). According to Lanke (1983:105):

> *... hot-deck imputation ... consists in substituting for each missing value a value determined in some way, deterministic or random, by values given by respondents who in some sense are similar to the non-responding ones. ... hot-deck imputation means that only information present in the sample under consideration is used when creating the values to be imputed.*

To illustrate the properties of hot-deck estimators, suppose for simplicity that the first $n_r < n$ (with $n - n_r = n_{nr}$) elements are labelled as respondents. Given simple random sampling without replacement, a hot-deck estimator of the population mean $\overline{Y}$ can be written in the form (Little & Rubin 1987:63):

$$\hat{\overline{Y}}_{HD} = \frac{1}{n}\left[n_r \overline{y}_r + n_{nr} \hat{\overline{y}}_{nr}\right] \qquad (6.9)$$

where $\hat{\overline{y}}_{nr}$ is the mean of the imputed values, namely:

$$\hat{\overline{y}}_{nr} = \frac{1}{n_{nr}} \sum_{i=1}^{n_r} L_i y_{r_i} \qquad (6.10)$$

and L_i is the number of times y_{r_i} is used as a donor value. Note that $\sum_{i=1}^{n_r} L_i = n_{nr}$, the number of missing values.

The properties of the hot-deck estimator of the population mean depend on the procedure used to generate the numbers L_1, L_2, ..., L_{nr} (Little & Rubin 1987:63). In section 6.2.6.1, two simple cases are discussed, namely where the imputed values are regarded as a probability sample selected with or without replacement from the recorded responses. Other variations of the basic technique of hot-deck imputation are also discussed, namely sequential hot-deck imputation, flexible matching imputation and distance function matching.

6.2.6.1. Randomised Hot-deck Imputation

In the randomised hot-deck imputation procedure, imputed values are obtained by means of a probability sample selected with or without replacement from the respondent values[4]. Consequently, randomised hot-deck imputation has the advantage of allowing classical design-based inferences to be made (Lessler & Kalsbeek 1992:216). If $r = \frac{n_r}{n} \geq 0{,}50$ imputation may be done *without replacement* but if $r < 0{,}50$, the hot-deck procedure must duplicate donor values *with replacement.*

6.2.6.1.1. With Replacement Duplication

Consider the randomised with replacement hot-deck estimator of the population mean, $\hat{\bar{Y}}^{(wr)}_{RHD} \equiv (6.9)$ in a single imputation class:

$$\hat{\bar{Y}}^{(wr)}_{RHD} = \frac{1}{n}\left[n_r \bar{y}_r + n_{nr} \hat{\bar{y}}_{nr}\right] \tag{6.11}$$

When evaluating $E(\hat{\bar{Y}}^{(wr)}_{RHD})$, two expectations are involved, namely E_1 (the expected value over the sampling distribution) and E_2 (the expected value over all possible imputations given the selected sample). Since:

$$E_2(\hat{\bar{Y}}^{(wr)}_{RHD} \mid n_r) = \bar{y}_r \tag{6.12}$$

the bias of this estimator is (Ford 1983:188):

$$\begin{aligned} Bias(\hat{\bar{Y}}^{(wr)}_{RHD}) &= E_1(\bar{y}_r) - \bar{Y} \\ &= Bias(\bar{y}_r) \end{aligned} \tag{6.13}$$

The randomised hot-deck estimator is therefore an unbiased estimator of the population mean if the respondent mean is also unbiased, i.e., if the missing values are random across the entire sample. Of course, in the case of H imputation classes, the bias is

---

[4] Randomisation can also be implemented by applying the sequential hot-deck to randomly sorted data (see section 6.2.6.2).

expressed in terms of the biases of the respondent means in each imputation class. As the imputation classes become more homogeneous, the size of $Bias(\bar{y}_{r_h})$ shrinks. Therefore, it is important for the reduction of bias to form imputation classes which are as homogeneous as possible with regard to reported and missing values. (Ford 1983:188.)

The variance of the estimator $\hat{\bar{Y}}_{RHD}^{(wr)}$ is given by Ford (1983:191) as:

$$V(\hat{\bar{Y}}_{RHD}^{(wr)} \mid n_r) = \frac{S^2}{n}\left[1+\left(\frac{n-n_r}{n_r}\right)\left(\frac{n+n_r-1}{n}\right)\right] \tag{6.14}$$

The variance (6.14) is greater than or equal to $V(\bar{y}_r \mid n_r) = V(\hat{\bar{Y}}_M \mid n_r)$ as given in (6.2). This can be seen by writing $V(\bar{y}_r \mid n_r)$, without the finite population correction, as:

$$V(\bar{y}_r \mid n_r) = \frac{S^2}{n_r} = \frac{S^2}{n}\left(1+\frac{n-n_r}{n_r}\right) \tag{6.15}$$

and noting in (6.14) that $\frac{n+n_r-1}{n} \geq 1$ (provided that $n_r \geq 1$) (Ford 1983:191). The randomised hot-deck imputation procedure may therefore lead to an estimator with a variance *larger* than that of the mean imputation estimator. Assuming simple random sampling and ignoring the finite population correction, the proportionate variance increase of $\hat{\bar{Y}}_{RHD}^{(wr)}$ over $\hat{\bar{Y}}_M = \bar{y}_r$ is at most 25% and this maximum is attained with a response rate of 50% (Little & Rubin 1987:64). For example, in a simple random sample with n = 1000 and a response rate of 50%:

$$\frac{V(\hat{\bar{Y}}_{RHD}^{(wr)} \mid n_r)}{V(\bar{y}_r \mid n_r)} = \frac{1+\frac{n_{nr}}{n_r}\left(\frac{n+n_r-1}{n}\right)}{1+\frac{n_{nr}}{n_r}} = \frac{1+\frac{500}{500}\left(\frac{1000+500-1}{1000}\right)}{1+\frac{500}{500}} = 1{,}2495$$

The variance (6.14) can be reduced by selecting donor values without replacement.

6.2.6.1.2. Without Replacement Duplication

Suppose simple random sampling without replacement is used to select donor values and consider again a single imputation class. The randomised without replacement hot-deck estimator $\hat{\bar{Y}}_{RHD}^{(wor)} \equiv (6.9) \equiv (6.11)$:

$$\hat{\bar{Y}}_{RHD}^{(wor)} = \frac{1}{n}\left[n_r\bar{y}_r + n_{nr}\hat{\bar{y}}_{nr}\right] \tag{6.16}$$

is an unbiased estimator of $\bar{Y}$ under the model of data missing at random across the entire sample. Its variance, ignoring the finite population correction, is given by Ford (1983:191) as:

$$V(\hat{\bar{Y}}_{RHD}^{(wor)} \mid n_r) = \frac{S^2}{n}\left[1 + \frac{2(n-n_r)}{n}\right] \tag{6.17}$$

Equation (6.17) is greater than or equal to $V(\bar{y}_r \mid n_r) = V(\hat{\bar{Y}}_M \mid n_r)$ but it is less than or equal to $V(\hat{\bar{Y}}_{RHD}^{(wr)} \mid n_r)$. Hence, selecting donor values *without* replacement may lead to a reduction in variance relative to selecting donor values *with* replacement. However, an estimator is still produced with variance greater than that of the class mean imputation estimator. The proportionate variance increase over $\bar{y}_r = \hat{\bar{Y}}_M$ is at most 12,5% and this maximum is attained when the response rate is 75% (Little & Rubin 1987:65). For example, in a simple random sample of size $n = 1000$ and a response rate of 75%:

$$\frac{V(\hat{\bar{Y}}_{RHD}^{(wor)} \mid n_r)}{V(\bar{y}_r \mid n_r)} = \frac{1 + \frac{2n_{nr}}{n}}{1 + \frac{n_{nr}}{n_r}} = \frac{1 + \frac{2 \times 250}{1000}}{1 + \frac{250}{750}} = 1{,}125$$

To summarise, both randomised hot-deck procedures (with and without replacement) yield larger variances than a procedure which imputes the respondent mean, but the randomised hot-deck variance can be reduced by duplicating donor values without replacement (Ford 1983:192). On the other hand, an advantage of the hot-deck procedure relative to mean imputation is that the distortion of the distribution of survey variable values in the sample is avoided. Furthermore, one should take into consideration that mean imputation actually gives an over-optimistic impression of the precision of estimators (a smaller variance is obtained compared with the variance from a sample consisting entirely of actually observed values).

6.2.6.2.Sequential Hot-deck Imputation

As with cold-deck imputation, most hot-deck imputation procedures begin by forming suitable imputation classes. In the sequential hot-deck imputation procedure, "seed" values of the item under consideration are assigned to each imputation class in the same way as for the cold-deck, i.e., responses are obtained from a previous survey which was conducted under similar survey conditions. These initial values are required in case the first value in an imputation class is missing. The data file is then ordered in preparation for a single pass required by the process. Each record of the current data file is considered sequentially one at a time. First, the record is associated with a specific imputation class. Then, if the record has a value for the item under consideration, that value replaces the (seed) value currently in the imputation class. If the response is missing, the (seed) value currently in that class becomes the imputed value for that variable. The process continues, replacing seed values with current responses where available and imputing seed or current class values to missing responses. (Lessler & Kalsbeek 1992:213.) Other than with the cold-deck method, the values stored in the hot-deck are continuously "updated".

Assuming only one imputation class, suppose the n sample values *and* the 1 seed value initially placed in the cold deck form a simple random sample of size $n + 1$ selected independently from the same population The sequential hot-deck estimator considered by Bailar, Bailey and Corby (1978) can be written as:

$$\hat{\bar{Y}}_{SHD} = \frac{1}{n}\sum_{i=0}^{n} L_i y_i \tag{6.18}$$

where y_0 is the seed value selected from the set of respondents and L_i is the number of times the value y_i is used, $L_i = 0$ if the i-th sample element is a non-respondent and $L_i \geq 1$ if the i-th sample element is the seed or a respondent. (Lessler & Kalsbeek 1992:216.)

The estimator $\hat{\bar{Y}}_{SHD}$ is unbiased under the model of data missing at random and its variance is given by Ford (1983:192) as:

$$V(\hat{\bar{Y}}_{SHD} \mid n_r) = \frac{S^2}{n}\left[1+\left(\frac{2(n-n_r)}{n}\right)\left(\frac{n_r n + n - 1}{(n_r+1)(n_r+2)}\right)\right] \tag{6.19}$$

The sum of the terms in square brackets represents the factor by which the variance of the sequential hot-deck estimator increases relative to the variance of the mean $\bar{y}$ in the case of complete response (ignoring the fpc). If $n_r > 0$, the variance of the sequential hot-deck estimator is also greater than the variance of the overall mean imputation estimator (the variance of the respondent mean) by a factor $\frac{n-n_r}{n}$, i.e., $V(\hat{\bar{Y}}_{SHD} \mid n_r) > V(\hat{\bar{Y}}_M \mid n_r) = V(\bar{y}_r \mid n_r)$.

For large n, $V(\hat{\bar{Y}}_{SHD} \mid n_r)$ can be approximated by:

$$AV(\hat{\bar{Y}}_{SHD} \mid n_r) = \frac{S^2}{n}\left[1+\left(\frac{2(n-n_r)}{n_r}\right)\right] \tag{6.20}$$

From (6.20) and (6.17) can be seen that $AV(\hat{\bar{Y}}_{SHD} \mid n_r) \geq V(\hat{\bar{Y}}^{(wor)}_{RHD} \mid n_r)$, since $n_r \leq n$. Thus, for large samples, the variance for a sequential hot-deck procedure is greater than or equal to the variance of a hot-deck procedure which duplicates randomly without replacement.

No estimator for $V(\hat{\bar{Y}}_{SHD} \mid n_r)$ has been suggested by authors on the sequential hot-deck procedure (Ford 1983:194). However, as before, one can replace S^2 in (6.19) by its unbiased estimator s_r^2 to obtain a variance estimator.

The sequential hot-deck procedure has a number of disadvantages. *Firstly*, the estimators are entirely dependent on the order of the file which may not be random; hence, the selection of a donor record may not be random. Without a probability mechanism governing the selection of donor values, it is impossible to calculate model-free estimates of the bias and variance of the estimator. On the other hand, some statisticians have made objections to randomising the order of the file (Ford 1983:196). They feel that in a non-random ordering, the possible correlation between nearby records may be exploited to improve the imputation (Sande 1982:149). For example, if the records are in geographic order, then duplicating the last response may be better than duplicating a random one - if there is a tendency for respondents who are geographically close to be similar in value (Ford 1983:196). The file may also be purposely ordered in a way that creates positive auto-correlation but there is then the danger of the multiple use of the same donor value if the ordering variable causes the missing values to cluster together (Ford 1983:196). Furthermore, as is often the case in practice, imputation classes are small and spread throughout the data file with the result that the benefit of correlation between neighbouring records is unlikely to be substantial (Kalton & Kasprzyk 1982:23).

Note that if the records in the data file are in random order, the sequential hot-deck method is equivalent to the selection of a simple random sample of donor values *with replacement* within each imputation class (ignoring the seed values used to start the process - see section 6.2.6.1) (Kalton & Kish 1981:146). Equation (6.19) should therefore be equal to equation (6.14) since the procedures are similar. Ford (1983:192) explains that the difference between the expressions arises because of the seed value selected from the population.

Bailar and Bailar (1978) have derived variance expressions for a procedure where the data file is not randomly ordered but is ordered so that there is serial correlation.

A *second* disadvantage of the sequential hot-deck is that the procedure may easily give rise to the multiple use of donors. For example, if within a given imputation class, a record with missing data is followed by a number of other records with missing data, all these records are assigned the current donor value in the imputation class, namely the seed value or the value of the last respondent. As is the case with mean imputation, the multiple use of some donor values contributes to lowering the precision of survey estimates and under-estimating variance estimators.

A *third* disadvantage of the sequential hot-deck is that the number of imputation classes has to be limited to ensure that one or more donors will be found for each class containing a record with a missing value. The danger of the multiple use of donor values is likely to increase when using a large number of imputation classes which so accurately divide the sample that some classes have many missing values but few responses. (Ford 1983:197.) Ford (1983:197) stresses that the auxiliary variables used to define imputation classes must be correlated with the missing-data item but not "too correlated with non-response". For example, if there is perfect correlation with non-response, then imputation classes will only have missing values and no responses that can be used as donor values.

Fourthly, because different items for the same record are imputed from different records, several donors may be involved in completing a single incomplete record; this may lead to inconsistencies among the imputed values (Kalton & Kish 1981:150). For example, suppose both the employment and education items are missing on a single record. In separate imputation, it may occur that a medical practitioner is "created" whose highest education level is Matric[5]. In order to avoid inconsistencies in imputed and reported values, it has been suggested that items which could lead to inconsistencies should have their responses deleted and also imputed for (Ernst 1978:471). For example, records with missing income values can have their employment responses deleted and imputed for. However, this is not a desirable solution to the problem.

Two advantages of the sequential hot-deck are (1) the imputations can be made from a single pass through the data file and (2) computationally, it is not a difficult procedure. For this reason, sequential hot-deck procedures have, in the past, been favoured for the processing of large data sets (Ford 1983:195).

6.2.6.3.Weighted Sequential Hot-deck Imputation

The sequential hot-deck procedure described in section 6.2.6.2 is "unweighted" because the selection of a donor record is independent of the sampling weight associated with donor and recipient records. However, ignoring sampling weights implies that the distribution of values within each imputation class of the clean data set may be distorted from that of the original distribution of responses. Cox (1980:721) developed a weighted sequential hot-deck algorithm that considers individual inclusion probabilities and uses probability selection of donors to minimise imputation bias. The algorithm is designed so that the expected values of estimators calculated from the clean data set will be equal to the weighted estimators using respondent data only (Cox 1980:721).

---

[5] This shows the importance of subjecting imputed values to the same edit checks as the actual responses to the survey (Lessler & Kalsbeek 1992:213).

The weighted sequential hot-deck has the additional advantages that (1) it controls the number of times that a donor record can be used for imputation and (2) it gives each respondent record a positive probability to be selected as a donor by applying a unique sequential sampling method to an ordered set of respondent data in the hot-deck (Lessler & Kalsbeek 1992:215).

A practical limitation of the weighted sequential hot-deck procedure is that multiple passes of the data file are required (Lessler & Kalsbeek 1992:215). The weighted sequential hot-deck algorithm and its properties are further described in Cox (1980).

6.2.6.4. Flexible Matching Imputation

The term "flexible matching imputation" is used to denote a sophisticated hot-deck procedure in which potential donors and recipients are matched on a sizeable number of auxiliary variables. The matching is done on a hierarchical basis, in the sense that if no donor can be found to match a recipient on all the auxiliary variables, some of the auxiliary variables considered less important are dropped or their detail of classification is reduced and the match is made at a lower level. With each failure at matching, imputation classes are further combined until a match can be found. Although the complexity of this matching process is obviously greater than in the sequential hot-deck procedure, this version has the advantage of adding flexibility to the matching so that the donor is as similar to the non-respondent as the data used for matching will allow. (Kalton 1983a:74.)

The flexible matching procedure enables far closer matches to be secured for many recipients than does the conventional hot-deck procedure. The procedure also avoids the multiple use of donors in classes where the number of donors is not less than the number of recipients.

According to Lessler and Kalsbeek (1992:215) empirical evaluation of flexible matching imputation seems to substantiate the usefulness of the technique.

6.2.6.5. Distance Function Matching

Imputation procedures which make use of imputation classes formed by the cross-classification of (categorical) auxiliary variables are generally not suitable for use with continuous auxiliary variables. Even if the continuous auxiliary variables are categorised, sequential hot-deck imputation may lead to some less desirable matches between recipients and donors. For example, a recipient with an auxiliary variable value close to the upper bound of the imputation class, may be matched with a donor close to the lower bound of the class, whereas it might have been better matched with a donor near the lower bound of the next higher class (Kalton 1983a:75). The use of a *distance function* to measure the closeness of a match between a recipient and a donor can be used to avoid such an occurrence.

In distance function matching, a donor value is defined in terms of a quantifiable measure of nearest distance to the recipient, where distance is measured as a function of the (continuous) auxiliary variables. Some statisticians consider distance function matching to be a "numerical" version of the hot-deck (Lessler & Kalsbeek 1992:218). Even when several auxiliary variables are used but only one is continuous, the technique of distance function matching can still be used. Imputation classes can first be formed from the categorical variables and the procedure can then be applied to the continuous variable within the imputation classes (Lessler & Kalsbeek 1992:219).

Suppose K auxiliary variables are used, and $x_{i1}, x_{i2}, \ldots, x_{iK}$ are the values of the K auxiliary variables for sample element i, where $i = 1, \ldots, n$. Define the distance between elements i and i' (Little & Rubin 1987:66):

$$d_1(i,i') = \max_k \left| x_{ik} - x_{i'k} \right| \tag{6.21}$$

The use of this distance function implies that the i-th and i'-th elements are considered only as similar as their most dissimilar difference among auxiliary variables (Lessler & Kalsbeek 1992:219). A donor value may then be chosen from respondent values such that $d(i,i')$ is less than some pre-determined value d_0 (Little & Rubin 1987:66).

Alternatively, if w_k measures the relative importance of the k-th auxiliary variable, the following distance measure can be used (Lessler & Kalsbeek 1992:218):

$$d_2(i,i') = \max_k w_k \left| x_{ik} - x_{i'k} \right| \tag{6.22}$$

The Mahalanobis distance measure has also been suggested for matching:

$$d_3(i,i') = \left(\underline{\mathbf{x}}_i - \underline{\mathbf{x}}_{i'} \right)' \sum \left(\underline{\mathbf{x}}_i - \underline{\mathbf{x}}_{i'} \right) \tag{6.23}$$

where $\sum$ is the estimated covariance matrix for the set of auxiliary variables and $\underline{\mathbf{x}}_i$ and $\underline{\mathbf{x}}_{i'}$ are respectively the $k \times 1$ vectors of auxiliary variable values for the i-th and $i' -$ th sample elements (Lessler & Kalsbeek 1992:219).

When this technique is used, attention must be paid to the distributions of the auxiliary variables. Distances in the tail of a distribution where the data are sparse are likely to dominate the overall distance function unless a suitable transformation is made. For example, economic data are often highly skewed toward the low end but the difference between R10 000 and R11 000 is R1 000 or 10%, while the difference between R100 000 and R105 000 is R5 000 or 5%. In some sense, the members of the second pair are more similar than the first. (Sande 1982:150.) For auxiliary variables with skewed distributions, distance functions are better formulated in terms of transformed variables such as $\log(x_{ik})$ or the rank of x_{ik} instead of the original x_{ik} (Lessler & Kalsbeek 1992:218).

According to Little and Rubin (1987:66) distance function matching requires considerable computing power and also an efficient search algorithm. However, the choice of distance function, given that the data have been suitably transformed, does not appear crucial and a distance measure that is computationally simple is advisable (Sande 1982:150).

As with the sequential hot-deck procedure, distance function matching allows donors to be used more than once. The chances of multiple uses can be controlled, however, by incorporating into the distance function the number of times the donor has previously been used so that the distance increases with the number of previous donations. Specifically, d can be multiplied by a factor such as $1+\lambda t$, where t is the number of times the donor has been used and λ is the assigned penalty for each usage. (Kalton 1983a:75). Other considerations, such as concern about the quality of measurement for a potential donor, may be reflected by adding another distance increment for the distance function (Sande 1982:150).

Since the selection of a donor value is deterministic and not probabilistic, the statistical properties of distance function matching are difficult to study without an explicit model. Sande (1982:150) notes in this regard that distance function matching can be converted into a hot-deck procedure by choosing the donor record at random from m nearest neighbours instead of taking the nearest satisfactory record.

Another variation of the distance function matching technique is to assign the recipient the average value for a set of nearest neighbours but this procedure suffers the disadvantage of distorting the empirical distribution of the sample (Kalton & Kasprzyk 1982:24).

6.2.7. Regression Imputation

Regression imputation uses values predicted by a regression of the missing item on a set of auxiliary variables. The functional form of the regression model is almost always linear (Lessler & Kalsbeek 1992:220), so that the i-th imputed value may be expressed as:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k X_{ki} + e_i \qquad (6.24)$$

where the β's can be estimated by standard methods such as ordinary least squares. Regression imputation differs from hot-deck methods in that the imputed value is a predicted value rather than an actual value taken from a designated respondent in the current survey.

The regression imputation method presumes that the item to be imputed is continuous and that the K auxiliary variables are all continuous, although qualitative variables may be used as well by defining appropriate indicator variables. Since regression imputation is basically a modelling technique (Little & Rubin 1987:61) it will not be discussed further in this chapter.

6.2.8. Multiple Imputations

All the imputation methods described above produce a single imputed value for each missing value. In general, this more or less distorts the natural distribution of values for that item (Särndal *et al.* 1992:594). Furthermore, standard variance formulas applied to the clean data set systematically under-estimate the variances of estimates even if the model used to generate the imputations is correct. As an improvement over single imputation methods, Rubin (1987) advocates the routine production of several sets of imputed values under different models or sets of assumptions as part of regular processing of survey results. This enables the statistician to (1) estimate the "imputation error" (that part of the error due to imputation) in the actual data so that the effects of different models can be studied (Sande 1982:151); (2) calculate valid estimates of the variance of estimates using standard complete data procedures[6] and (3) alleviate, to some extent, the under-estimation of variability due to single imputation methods (see sections 6.2.3 and 6.2.6.2).

Suppose for example, that I distinct imputations are made for each of the n_{nr} missing values for a single item, to form I complete data sets. From these I imputed data sets, I complete data statistics are calculated. For example, the mean of each complete data set $\hat{\bar{y}}_1, \hat{\bar{y}}_2, \ldots, \hat{\bar{y}}_I$ can be calculated as:

$$\hat{\bar{y}}_i = \frac{1}{n}\left[n_r \bar{y}_r + n_{nr} \hat{\bar{y}}_{nri}\right] \tag{6.25}$$

for $i = 1, \ldots, I$ where $\hat{\bar{y}}_{nri}$ is the mean of the n_{nr} imputed values in data set i. And the estimated variance $\hat{s}_1^{\,2}, \hat{s}_2^{\,2}, \ldots, \hat{s}_I^{\,2}$ of each complete data set can be calculated as:

$$\hat{s}_i^{\,2} = \frac{1}{n-1}\left[(n_r - 1)s_r^{\,2} + n_r \bar{y}_r^{\,2} + n_{nr}(\hat{\bar{y}}_{nri})^2 - n(\hat{\bar{y}}_i)^2\right] \tag{6.26}$$

for $i = 1, \ldots, I$ (Hertzog & Rubin 1983:215).

From the I estimates, a single combined estimate of the population mean is computed, namely the mean of the I complete data set means:

$$\hat{\bar{Y}}^{*} = \frac{1}{I}\sum_{i=1}^{I} \hat{\bar{y}}_i \tag{6.27}$$

A pooled variance estimate is computed with two variance components: the one reflecting the average variance within the imputations and the other the variance across the imputations of the I means. The latter component can be seen as an expression of the error due to imputation (Särndal *et al.* 1992:594).

Hertzog and Rubin (1983:217) give the average variance of the I imputed data sets as:

---

[6] By "standard complete data methods" is meant calculating the mean and variance in the usual way, as if it were a complete matrix of actual responses.

$$\frac{\overline{\hat{s}^2}}{n} = \frac{1}{I}\sum_{i=1}^{I}\frac{\hat{s}_i^{\,2}}{n} \tag{6.28}$$

and the variance across imputations of the complete data set means (i.e., the average variance of the mean estimators of each imputed data set) as:

$$W = \frac{1}{I-1}\sum_{i=1}^{I}\left(\hat{\bar{y}}_i - \hat{\bar{Y}}^{\bullet}\right)^2 \tag{6.29}$$

Hertzog and Rubin (1983:218) show that since:

$$E(\hat{\bar{Y}}^{\bullet}) = \bar{y}_r \tag{6.30}$$

the multiple imputation estimator is unbiased under the model. Furthermore:

$$V^{\bullet}(\hat{\bar{Y}}^{\bullet} \mid n_r) = \frac{n-n_r}{n^2}\frac{s_r^{\,2}}{I} \tag{6.31}$$

Hertzog and Rubin (1983:218) derive an expression for the unconditional variance (over repeated sampling and imputation procedures) and show that there is a "real" reduction in the variance of the estimator of $\bar{Y}$ when using multiple random imputations rather than a single random imputation.

There are two improvements due to multiple random imputations over single random imputation: (1) the real variance of estimation is reduced and (2) the under-estimation of variability that follows from performing a single imputation is partially adjusted for. On the other hand, a disadvantage of multiple imputations is that it requires more work for data handling and computation of estimates (Särndal *et al.* 1992:594).

6.3. CONCLUSION

A major attraction of imputation is that it generates a complete data set that may readily be used for many different forms of analyses (Kalton & Kasprzyk 1982:29). Of course, an important aim of imputation is also to reduce the bias of non-response. However, there can be no guarantee that the results obtained after imputation will be less biased than those based on the incomplete data set and, indeed, the biases could be greater (Kalton & Kasprzyk 1982:22). The success of the imputation procedure depends on the validity of the model assumptions, i.e., bias will be reduced to the extent that the data are truly missing at random across the entire sample or to the extent that imputation classes can be formed so that data are missing at random within each imputation class.

It should be stressed that imputation can never create any *new* information as a substitute for the missing information. The main aim of an imputation procedure is to produce a clean data set in order to simplify data processing. (Lanke 1983:105.) Most authors warn of this danger of imputation, namely, that researchers may falsely treat the completed data set as if it were from a straightforward sample of size n. According to Dempster and Rubin (1983:8), as:

> *The idea of imputation is both seductive and dangerous. it is seductive, because it can lull the user into the pleasurable state of believing that that data are complete after all, and it is dangerous because it lumps together situations whether the problem is sufficiently minor that it can legitimately be handled this way and situations where standard estimators applied to real and imputed data have substantial biases.*

Analysts working with a data set that contains imputed values should be fully aware of the extent of imputation as well as the details of the procedures used. The imputation process should therefore be carefully monitored and recorded (Sande 1982:151). The use of imputation may give rise to ethical problems if not only the survey estimates but also the micro-data are going to be published. At the very least, all imputed values should be "flagged" to distinguish them from the actual responses (Särndal *et al.* 1992:591; Kalton & Kasprzyk 1982:22). In some countries, the presence of imputed values in data files on individuals are forbidden by law, even if such values are flagged to indicate their artificial origin (Särndal *et al.* 1992:591). Sande (1982:151, 1983:346) recommends that at least the following information should be made known together with the survey results:

1. The number of records in which any imputation is made
2. The number of records missing specific variables (or combinations of variables) due to item non-response and those due to edit failure
3. The number requiring one (two, three etc.) item(s) to be imputed
4. How many times a particular record has been used as a donor[7]
5. How many attempts were required to complete a particular record

[7] There does not seem to be a generally accepted guideline as to the minimum response rate required before imputations can be made or the number of times a particular donor value can be used.

6. Which donors contributed what fields to which recipients
7. What the value of the distance function was (d_0)
8. A listing of any records failing to be completed

Monitoring the imputation procedure (and equally important, the editing procedure that precedes the imputation) can give information about the effectiveness of the edits and the imputation procedure, leading to improvements in subsequent versions of the survey or even in other surveys (Sande 1982:151).

An important consideration in the choice of imputation method that has not been mentioned yet is *the type of variable being imputed for*. All the imputation methods discussed can be applied with continuous variables but some of them are not suitable for use with categorical or discrete variables. For example, a class mean imputation value of 10,7 for the discrete variable "number of children in a household" or a regression imputation value of 0,7 for the nominal variable "marital status" are not feasible for individual respondents and rounding them to whole numbers may lead to bias. For this reason, these imputation methods do not work well for categorical (nominal or discrete) variables. On the other hand, hot-deck methods always give feasible values since the values are taken from actual respondents. (Kalton & Kasprzyk 1986:8).

It should be noted that weighting adjustments (to compensate for unit non-response) and imputation (to adjust for item non-response) are usually employed in combination. Imputation is therefore rarely applied to a self-weighting sample. Nevertheless, this chapter has concentrated on the implementation of imputation procedures to self-weighting samples. Kalton and Kasprzyk (1982:30) mention that more research is needed on the implications of unequal weights on imputation (see section 6.2.6.3).

Lastly, it should be recognised that the typical survey collects data on a substantial number of variables, often up to a hundred variables or more. Since all variables are likely to have some missing responses, the task of creating a complete data set is formidable. It is generally not practicable to develop a separate tailor-made imputation method for each variable. At best, a separate procedure may be developed for only a small selection of the most important survey variables. (Kalton & Kasprzyk 1982:28.)

CONCLUSION

It is about 12 years now since I finally came to the sad conclusion that most of the statistical methods that I had learned from pioneers like Karl Pearson, Ronald Fisher, and Jerzey Neyman and survey practitioners like Morris Hansen, P.C. Mahalanobis, and Frank Yates are logically untenable. It is my interest in survey theory that finally forced me to this unhappy conclusion.

Debabrata Basu, A Discussion of Survey Theory

It is true that the perfect (probability) survey (as described in the introduction to this dissertation) is a myth - an ideal which can only be attained in the clinical milieu of the textbook. There are innumerable sources of error that detract from the ideal. In this dissertation, the one source of error was discussed that has captured the attention of many practitioners; probably because the non-response rate can so easily be calculated and documented. In fact, the non-response rate is often used (incorrectly so) as an indicator of the quality of survey data. As a consequence, the total attempt to deal with the non-response problem in many surveys amounts to the calculation of some non-response rate. In many survey reports, the non-response rate is mentioned in a footnote (and forgotten) and the available (respondent) data are treated as a probability sample from the population. Statistical inferences are made to the entire population instead of to the responding sub-set only. In **Part I** of this dissertation (Chapters 1 and 2), the dangers of such an approach (or lack of approach) to the problem of non-response were described.

Chapter 1 prepared the setting for a discussion of the non-response problem. This chapter consisted of three main sections with the aim to:

1. Define the various terms, concepts and notation used throughout the dissertation
2. Discuss the sampling strategies that are relevant in the dissertation
3. Describe the approach to statistical inference that is assumed in the dissertation

The following three points emerged from this introductory chapter:

a) The non-response literature consists of contributions from a wide variety of researchers from different fields. Although the non-response methodology certainly benefits from the variety of perspectives and approaches introduced by the different disciplines, a study of the non-response problem is complicated by the fact that each field of research has its own vocabulary. There is a definite need for standardising the non-response vocabulary.

b) Most of the methods of dealing with non-response assume simple random sampling. There is a need to develop the theory for dealing with non-response in the more complex sampling designs that are used in "real-world" surveys.

c) In recent years, theoretical statisticians have focused increasingly on the use of *models* in statistical inferences. On the other hand, the modelling approach has not been widely adopted by survey practitioners who prefer the robustness of traditional *design-based* inferences. However, when it comes to the treatment of non-sampling errors such as non-response, the design-base must be supplemented by certain model-assumptions, for example, assumptions about the distribution of responses in the population. The resulting approach (which was followed in Part II of the dissertation) is called the *model-assisted design-based* approach. A drawback of this approach is that the validity of statistical inferences depends on the truth of assumptions which cannot usually be verified from the survey data themselves.

Chapter 2 of the dissertation consisted of four main sections with the aim to:

1. Identify and categorise the many diverse reasons why people do not respond
2. Examine the effect of non-response on survey estimates and derive an expression for the bias of non-response
3. Systematise the many different ways of calculating the response rate
4. Discuss the interpretation of a survey response rate

A number of important conclusions and recommendations arise from Chapter 2:

a) In any particular survey, the reasons for non-response should be identified and recorded according to or based on the categories identified in section 2.2. This may provide valuable information that can be used to:

   i) assess the impact of non-response on survey estimates

   ii) calculate response and completion rates and

   iii) devise a strategy for dealing with non-response[1].

b) The importance of constructing a good sampling frame cannot be over-emphasised. Besides the obvious advantage of reducing coverage error, a large proportion of non-response due to "non-contact" can be avoided by the use of sampling frames containing up-to-date addresses, telephone numbers or other auxiliary information that can be used to locate sample elements.

c) The proportion of survey non-response is often mistakenly seen as the sole indicator of the quality of survey estimates. However, the effect of non-response on survey estimates depends jointly on (i) the magnitude of the non-response rate and (ii) the differences between respondents and non-respondents in the population. It is important that efforts are made in a particular survey to examine the nature of the differences between respondents and non-respondents

[1] For example, in some surveys the impact of non-response due to *refusal* may be more severe than the impact of non-response due to *not-at-home* (and may require separate treatment).

even if a large response rate is obtained: *a large response rate does not necessarily guarantee negligible bias*! (Unfortunately, differences between respondents and non-respondents can usually not be evaluated from the survey data themselves - it often requires special efforts as described in Chapter 3.)

d) It is important that the survey response rate is based (i) *only* on eligible survey elements and (ii) on *all* eligible survey elements. The exclusion of certain elements, for example, those that could not be found, will result in under-estimating the true level of non-response in the survey. Similarly, the inclusion of non-eligible elements will result in over-estimating the non-response rate. However, eligibility status can often not be determined for all sample elements.

e) Whichever definition of the response rate is used in a particular survey, it is important that both the numerator and the denominator of the response rate be explicitly defined. The preparation of one or more accountability tables is recommended (see section 2.5.4). Accountability tables should be prepared for the entire sample and for important domains and should be included in the survey report. This will allow response and completion rates to be defined in terms of entries in the accountability tables and will facilitate comparisons on a common basis between different surveys, countries and survey organisations.

f) Almost every survey should be planned assuming that non-response will occur (Madow, Nisselson & Olkin 1983:13). Informed guesses about expected response rates and biases are necessary in the planning phase of the survey, for example, to define weighting or imputation classes, to determine the optimum number of calls to make on each sample element or to determine the optimum sub-sampling fraction to follow-up. Speculations about response rates and biases (for the survey as a whole and for important items and domains) should be based on previous experience or similar surveys. For this purpose, it is recommended that a systematic summarisation of information from various local surveys be undertaken, including information on response rates for specified types of populations and for particular questions in stated contexts (Madow, Nisselson & Olkin 1983:13). Although such information exists to a large extent for overseas populations, it is lacking for surveys conducted in the RSA.

In **Part II** of this dissertation (Chapters 3 to 6), an attempt was made to provide a comprehensive list of important model-assisted design-based methods that are currently available to deal with non-response. In general, one can do two things to compensate for missing data, namely:

1. Seek to obtain more complete data and
2. Modify the estimation of population values and analysis of the survey data

The first technique, i.e., employing *preventive* strategies, was covered in Chapter 4; the second, i.e., applying *post hoc* adjustment procedures to compensate for non-response, was covered in Chapters 5 and 6.

Whichever method is used to deal with non-response, one should also attempt to determine *empirically* what the extent of the damage is by estimating non-response biases during and after data collection and before and after adjustments for non-response[2]. Although this is a desirable thing to do, it is usually very difficult: recall from Chapter 2 that one component of non-response bias is the average difference between respondents and non-respondents in the population. To obtain *quantitative* estimates of this component from the survey data is usually impossible, since no data on the survey variables are available for the non-respondents - simply because they did not respond. However, there are at least two techniques that can be used in some surveys to estimate (usually *qualitatively*) the extent of respondent/non-respondent differences and, hence, the extent of non-response bias. These techniques, which were discussed in **Chapter 3**, are:

1. *The identification of auxiliary variables whose values are available for both respondents and non-respondents and which are significantly correlated with the major survey variables and with response behaviour.*

   The identification of such auxiliary variables is useful because respondent/non-respondent differences in auxiliary variable values usually indicate respondent/ non-respondent differences in the values of the survey variables. Auxiliary variable values may be obtained from previous surveys, administrative sources or other records. When designing the survey questionnaire, it is useful to include questions that are related to items with high expected non-response rates. These questions may be simpler or less threatening versions of sensitive questions, for example, questions on rent or housing expenditure in the case of non-response on income questions (Madow, Nisselson & Olkin 1983:9).

   Various techniques, for example, the box plot, CHAID-analysis and logistic regression can be used to identify suitable auxiliary variables[3].

2. *The analysis of respondent characteristics at various stages (call-attempts or waves) in the data collection process.*

   The assumption behind this procedure is that late responders indicate greater reluctance to participate and hence, stronger resemblance to non-respondents. The difference between late and early responders may be used to estimate differences between respondents and hard-core non-respondents. A regression model is sometimes used to extrapolate the cumulative responses obtained on successive call-attempts or waves to a 100% response rate. Such extrapolation

[2] One would indeed be fortunate to be able to prove in a particular survey that the non-response is *ignorable*, i.e., that the respondents are a random sub-sample from the population.

[3] Further utilisation of auxiliary variables identified by means of these techniques is in the formation of weighting and imputation classes for non-response adjustments (Chapters 5 and 6).

methods (and most other methods of trying to obtain quantitative estimates of non-response bias) have not been very successful.

Chapter 3 concluded with a discussion of various studies which have aimed to analyse differences between respondents and non-respondents in specific surveys. These (and many other) studies have established that there are usually important differences between respondents and non-respondents in many surveys and that these differences are usually related to the survey variables of interest. For example, respondents are usually found to have higher incomes and higher educational levels than refusers. However, these findings apply mainly to populations in Europe and the USA. Research should be conducted to determine to what extent respondents differ from non-respondents in South African surveys.

Chapter 4 considered various techniques that can be used to prevent a high non-response rate. This chapter consisted of four main sections:

1. A discussion of various operational procedures and principles of survey and questionnaire design that aim to increase the likelihood of response for all sample elements
2. A consideration of the practice of substituting each non-respondent with a randomly selected or specially designated substitute or selecting more than the required number of elements to allow for "shrinkage"
3. A discussion of Deming's model that allows characteristics of the population and fieldwork costs to be incorporated into determining the optimum number of call-backs to be made
4. A discussion of a technique by Hansen and Hurwitz and an alternative technique by Srinath for determining the optimum sub-sampling fraction and the initial sample size which minimise the expected cost of the survey for a desired precision of the estimator

The following can be noted from Chapter 4:

a) A large amount of research has been conducted (mostly in the social sciences) on techniques such as advance notification, personalisation, questionnaire length, size and colour, as well as financial and material incentives. Although no single "magic bullet" can be identified that will significantly stimulate response, it is true that reasonably high survey response rates can be obtained by using and integrating multiple techniques.

b) Substitution is a method that should be labelled: "Handle With Care". *Firstly*, researchers may (in the words of Oh and Scheuren (1983:145)) be "lulled into believing" that a substitute is just as good as the originally selected element and, consequently, use less effort to obtain responses. *Secondly*, substitution may not reduce non-response bias at all, since substitutes usually resemble respondents (and not non-respondents) simply because they *responded*. *Thirdly*, substitutes are usually mistakenly treated as respondents when calculating response rates. The effectiveness of substitution to reduce non-response bias depends on the (unknown) degree that each substitute resembles the sample element it is replacing.

c) The application of Deming's model for determining the optimum number of call-backs to be made requires prior knowledge of the response probabilities in the population, the expected response rates per call, the relative costs per interview, the estimated biases per call and the population variances. Reasonable estimates of these values may be obtained from analyses of call-back data in past surveys with the same types of questions and the same populations. For this reason, it is recommended that information is accumulated about costs, relative biases, response rates, etc., in South African surveys. This will allow an economic call-back policy to be worked out in advance for any type of survey. (This can be identified as an area which still requires a considerable amount of research!)

d) It is usually impractical and costly to call back repeatedly until complete responses are obtained from all non-respondents. However, unbiased estimation is still possible if a 100% response rate can be obtained in a random *sub-sample* of all non-respondents. The Hansen and Hurwitz procedure for determining the optimum sampling fraction in the sub-sample requires prior knowledge of expected costs in the initial and follow-up calls, as well as the expected variances among respondents and non-respondents and the expected population non-response rate. If the sub-sampling fraction and the response rate in the sub-sample are high, then sub-sampling may reduce the risk of non-response bias sufficiently to compensate for the increase in variance associated with the reduction in sample size due to interviewing only a sub-sample of non-respondents in the follow-up. But in general, the Hansen-Hurwitz procedure is useful only if the expected costs in the sub-sample are very large in relation to the costs of the initial call. This may be the case, for example, when mail questionnaires are followed up by relatively more expensive face-to-face interviews.

Even if all the techniques in Chapter 4 are applied and much effort is expended to collect data as fully as possible, some residual non-response is inevitable. This necessitates the consideration of methods that statistically *adjust* the collected data during the estimation and analysis phase of the survey. According to Dempster and Rubin (1983:7), both preventive strategies and post hoc adjustment procedures will only be partially successful:

> *Analysis is essential since the best collection efforts will leave residual incompleteness, often substantial residual incompleteness. Nevertheless, when data are incomplete, the performance of any data collection scheme and analysis procedure is invariably somewhat reliant on unverifiable assumptions.*

Two different methods are generally used to compensate for the two types of non-response: unit non-response is usually compensated for by increasing the *weights* of specified respondents, while item non-response is usually compensated for by means of some *imputation* procedure.

In **Chapter 5**, various techniques were considered that adjust the weights of respondents in the sample. This chapter consisted of four main sections:

1. A discussion of the (mainly theoretical) technique of increasing the weights of all respondents by a constant amount
2. A discussion of the technique of increasing the weights of respondents within certain sub-groups of the sample when (a) the population sizes of the sub-groups are unknown and (b) the population sizes of the sub-groups are known
3. A discussion of the technique of raking ratio estimation to adjust for non-response
4. A discussion of various other weighting techniques such as sub-sample weighting for non-response, the Politz-Simmons procedure and linear regression estimation

The most significant conclusions in Chapter 5 can be summarised as follows:

a) Weighting all respondents inversely proportional to the sample response rate, implies the assumption that the data are missing at random across the entire sample. This basically means that non-response is assumed ignorable. However, as was shown in Chapter 3, non-respondents are seldom a random sub-sample from the population. This weighting procedure therefore has little practical value.

b) An improvement over the assumption of data missing at random globally is to assume that there is random missing data within certain sub-groups (called weighting classes) of the sample. In this case, respondents within each weighting class are weighted by a constant factor. The ability of any sub-group weighting procedure to reduce non-response bias depends on the formation of weighting classes in which the assumption of uniform response probabilities is tenable. In fact, non-response bias may even be increased if inappropriate weighting classes are formed.

c) The technique of raking ratio estimation can be used to adjust for non-response. In this case, weighting classes are formed by cross-classifying two or more categorical auxiliary variables whose population marginals are known. The same requirements for bias reduction apply as in the case of sample and population weighting adjustments.

In **Chapter 6**, five different types of imputation methods for item non-response were discussed, namely:

1. Deductive imputation
2. Mean imputation
3. Cold-deck imputation
4. Hot-deck imputation
5. Regression imputation

A number of hot-deck imputation methods were discussed, namely:

a) Randomised hot-deck imputation (with and without replacement)
b) Weighted and unweighted sequential hot-deck imputation
c) Flexible matching imputation
d) Distance function matching

Clearly, practitioners have a large number of imputation methods to choose from. As a first step, one should check the responses to other questions to determine if logical, *deductive* imputations can be made for the missing data. *Mean* imputation and *cold-deck* imputation have certain disadvantages that have made them less popular to use. *Unweighted sequential hot-deck* imputation is easiest to use but is inappropriate when dealing with quantitative variables or when the response rate is low. *Weighted sequential hot-deck* imputation is more difficult to implement but is ideal when response rates are low and good predictive variables are not available for model building (Cox & Cohen 1985:235). *Regression* imputation is useful when dealing with quantitative variables for which a good predictive regression equation can be developed.

According to Oh and Scheuren (1983:145), much of current practice is oriented to applying adjustments for non-response with as much skill as possible and then acting as though the adjustments are adequate, with little, if any, discussion of either the assumptions or possible resulting biases. They warn (Oh & Scheuren 1983:145):

> *There are, of course, disadvantages in modeling the response mechanism as an additional stage of sampling. One of these is that some practitioners may be lulled into the belief that their results have the same robustness as probability sampling inferences in the complete data case - a misconception that could have particularly disastrous consequences if there is a high nonresponse rate.*

It is important that practitioners who use weighting or imputation bear in mind that the success of their adjustments to reduce non-response bias depends on two factors; namely (Oh & Scheuren 1983:145):

1. Whether or not the response probabilities are truly positive for all elements
2. Whether or not the relationship between the probabilities of a response and the auxiliary variables used to define weighting or imputation classes has been modelled correctly

Unfortunately, (1) there are almost always hard-core unit or item non-respondents in a survey and (2) no matter how well the survey is designed, it is impossible to observe all the characteristics that determine the probability of a response. Nevertheless, it is important that practitioners (a) explicitly state the response mechanism assumed by the adjustment procedure they employ, (b) cite evidence, if available, on why such a mechanism may be plausible and (c) perform analyses on the sensitivity of inferences to alternative specifications of the response mechanism (Oh and Scheuren 1983:158).

An important aim of this dissertation was to answer the question: "What can be done about the error of non-response?" A large number of techniques were discussed that *can* be successfully used to reduce or limit non-response bias. According to Madow, Nisselson and Olkin (1983:6):

> *The general methods currently most likely to reduce biases are those employing poststratification. The methods may utilize imputation or weighting techniques.* [But] *In general, no statistical methodology for imputation or adjustment will reduce the need to attempt to collect data with high levels of response.*

According to Oh and Scheuren (1983:144):

> *The models employed in adjusting for missing data, no matter how cleverly structured, virtually never hold in practice; hence, the more non-response present, the greater is the sensitivity of one's results to the mechanism assumed in carrying out the adjustments.*

Oh and Scheuren (1983:181) also state:

> *Since response models almost never hold exactly, the only truly robust approach to the problem of bias is to keep nonresponse to a minimum. There is no adequate substitute for complete response ...*

In the light of the above, it is important to note that:

> **There is no totally satisfactory substitute for complete or nearly complete response. As a consequence, the best approach to dealing with non-response is to collect survey data as fully and accurately as possible.**

At the conclusion of this dissertation, one must recognise that:

1. An "ignorance is bliss" attitude is inexcusable - there is a plethora of tools and techniques available that can be used to deal with non-response.

However, after having applied one or more of these techniques, researchers should ardently guard against being lulled into the pleasurable state of believing that the non-response problem has been done away with since, ultimately:

2. The problem of non-response bias is never completely resolved.

Madow, Nisselson and Olkin (1983:6) agree:

> *... no statistical methods will fully compensate for missing units and data. Biases will almost certainly remain. Good methods are aimed chiefly at reducing biases and mean square errors of estimators while reducing or at least not unduly increasing variances of estimators.*

SOURCES CONSULTED

Adams, L.L. & Gale, D. 1982. Solving the Quandary between Questionnaire Length and Response Rate in Educational Research. *Research in Higher Education*, 17(3):231-240.

Alho, J.M. 1990. Adjusting for Nonresponse Bias Using Logistic Regression. *Biometrika*, 77(3):617-624.

Anderson, H. 1979. On Nonresponse Bias and Response Probabilities. *Scandinavian Journal of Statistics*, 6:107-112.

Andrews, F.M. 1984. Construct Validity and Error Components of Survey Measure. *Public Opinion Quarterly*, 48:409-442.

Armstrong, J.S. & Overton, T.S. 1977. Estimating Nonresponse Bias in Mail Surveys. *Journal of Marketing Research*, 14:396-402.

Ayidiya, S.A. & McClendon, M.J. 1990. Response Effects in Mail Surveys. *Public Opinion Quarterly*, 54:229-247.

Babbie, E. 1990. *Survey Research Methods*. 2nd edition. Belmont: Wadsworth.

Bailar, B.A. 1976. Some Sources of Error and Their Effect on Census Statistics. *Demography*, 13:273-286.

Bailar, B.A. 1983. Counting or Estimation in a Census - A Difficult Decision, in *American Statistical Association Proceedings of the Social Statistics Section*. Washington, D.C.: American Statistical Association:42-47.

Bailar, B.A. & Bailar, J.C. 1983. Comparison of the Biases of the Hot-deck Imputation Procedure with an "Equal-Weights" Imputation Procedure, in *Incomplete Data in Sample Surveys Volume 3: Proceedings of the Symposium*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press: 299-311.

Bailar, B.A., Bailey, L. & Corby, C. 1978. A Comparison of Some Adjustment and Weighting Procedures for Survey Data, in *Survey Sampling and Measurement*, editor N.K. Namboodiri. New York: Academic Press:175-198.

Bailar, J.C. & Bailar, B.A. 1978. Comparison of Two Procedures for Imputing Missing Survey Values, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:462-467.

Bailey, K.D. 1987. *Methods of Social Research*. 3rd edition. New York: Free Press.

Bailey, L. 1983. Compensation for Unit Nonresponse in Recurring Surveys, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:289-294.

Bankier, M.D. 1986. Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys. *Journal of the American Statistical Association*, 81:1074-1079.

Banks, M.J. 1977. An Indication of the Effect of Noninterview Adjustment and Post-stratification on Estimates from a Sample Survey, in *American Statistical*

Association Proceedings of the Social Statistics Section. Washington, D.C.: American Statistical Association:291-295.

Bartholomew, D.J. 1961. A Method of Allowing for "Not-at-Home" Bias in Sample Surveys. *Applied Statistics*, 10:52-59.

Basu, D. 1983. A Discussion of Survey Theory, in *Incomplete Data in Sample Surveys Volume 3: Proceedings of the Symposium*, editors W.G. Madow & I. Olkin. New York: Academic Press: 407-410.

Bethlehem, J.G. 1988. Reduction of Nonresponse Bias Through Regression Estimation. *Journal of Official Statistics*, 4:251-260.

Bethlehem, J.G. & Keller, W.J. 1987. Linear Weighting of Sample Survey Data. *Journal of Official Statistics*, 3:141-154.

Bethlehem, J.G. & Kersten, H.M.P. 1981a. *Graphical Methods in Non-response Analysis and Sample Estimation.* The Hague: Staatsuitgeverij. (Statistical Studies; no. 31.)

Bethlehem, J.G. & Kersten, H.M.P. 1981b. The Nonresponse Problem. *Survey Methodology*, 7(2):130-156.

Bethlehem, J.G. & Kersten, H.M.P. 1985. On the Treatment of Nonresponse in Sample Surveys. *Journal of Official Statistics*, 1:287-300.

Birnbaum, Z.W. & Sirken, M.G. 1950. Bias Due to Nonavailability in Sampling Surveys. *Journal of the American Statistical Association*, 45:98-111.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. 1975. *Discrete Multivariate Analysis: Theory and Practice.* Massachusetts: MIT Press.

Brackstone, G.J. & Rao, J.N.K. 1981. An Investigation of Raking Ratio Estimators. *Sankhyā : The Indian Journal of Statistics*, Series C, 41:97-114.

Brick, J.M., Waksberg, J., Kulp, D. & Starer, A. 1995. Bias in List-assisted Telephone Surveys. *Public Opinion Quarterly*, 50(2):218-235.

Brunner, G.A. & Carroll, S.J. 1967. The Effect of Prior Telephone Appointment on Completion Rates and Response Content. *Public Opinion Quarterly*, 31:652-654.

Brunner, G.A. & Carroll, S.J. 1969. The Effect of Prior Notification on the Refusal Rate in Fixed Address Surveys. *Journal of Advertising Research*, 9:42-45.

Buck, S.F. 1960. A Method of Estimating Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. *Journal of the Royal Statistical Society*, Series B, 22:302-306.

Cassel, C.M., Särndal, C.E. & Wretman, J.H. 1977. *Foundations of Inference in Survey Sampling.* New York: Wiley.

Chadwick, B.A., Bahr, H.M. & Albrecht, S.L. 1984. *Social Science Research Methods.* Englewood Cliffs: Prentice Hall.

Chalmers, A.F. 1982. What is this Thing Called Science? An Assessment of the Nature and Status of Science and its Methods. Milton Keynes: Open University Press.

Chapman, D.W. 1976. A Survey of Nonresponse Imputation Procedures, in *American Statistical Association Proceedings of the Social Statistics Section.* Washington, D.C.: American Statistical Association:245-251.

Chapman, D.W. 1983. The Impact of Substitution on Survey Estimates, in *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press:45-61.

Chapman, D.W., Bailey, L. & Kasprzyk, D. 1986. Nonresponse Adjustment Procedures at the U.S. Bureau of the Census. *Survey Methodology*, 12:161-179.

Childers, T.L. & Ferrell, O.C. 1979. Response Rates and Perceived Questionnaire Length in Mail Surveys. *Journal of Marketing Research*, 16:429-431.

Chromy, J.R. & Horvitz, D.G. 1978. The Use of Monetary Incentives in National Assessment Household Surveys. *Journal of the American Statistical Association*, 73:473-478.

Claussen, J.A. & Ford, R.N. 1947. Controlling Bias in Mail Questionnaires. *Journal of the American Statistical Association*, 42:497-511.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd edition. New York: Wiley.

Cochran, W.G. 1983. Historical Perspective, in *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press:11-25.

Colledge, M.J., Johnson, J.H., Paré, R. & Sande, I.G. 1978. Large Scale Imputation of Survey Data, in *American Statistical Association Proceedings of the Section on Survey Research Methods.* Washington, D.C: American Statistical Association:421-436.

Cox, B.G. 1980. The Weighted Sequential Hot-deck Imputation Procedure, in *American Statistical Association Proceedings of the Section on Survey Research.* Washington, D.C.: American Statistical Association:721-726.

Cox, B.G. & Cohen, S.B. 1985. *Methodological Issues for Health Care Surveys.* New York: Dekker.

Cox, B.G. & Folsom, R.E. 1978. An Empirical Investigation of Alternative Nonresponse Adjustments, in *American Statistical Association Proceedings of the Section on Survey Research Methods.* Washington, D.C.: American Statistical Association:219-221.

Cox, B.G. & Folsom, R.E. 1981. An Evaluation of Weighted Hot-deck Imputation for Unreported Health Care Visits, in *American Statistical Association Proceedings of the Section on Survey Research Methods.* Washington, D.C.: American Statistical Association:412-417.

Cox, D.R. & Hinkley, D.V. 1974. *Theoretical Statistics.* London: Chapman and Hall.

Dalenius, T. 1961. Treatment of the Non-response Problem. *Journal of Advertising Research*, 1(5):1-7.

Daniel, W.W. 1975. Nonresponse in Sociological Surveys: A Review of Some Methods for Handling the Problem. *Sociological Methods and Research*, 3:291-307.

David, M., Little, R.J.A., Samuhel, M.E. & Triest, R.K. 1986. Alternative Methods for CPS Income Imputation. *Journal of the American Statistical Association*, 81:29-41.

Deming, W.E. 1944. On Errors in Surveys. *American Sociological Review*, 9:359-369.

Deming, W.E. 1950. *Some Theory of Sampling*. New York: Wiley.

Deming, W.E. 1953. On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Response and the Bias of Nonresponse. *Journal of the American Statistical Association*, 48:743-772.

Dempster, A.P. & Rubin, D.B. 1983. Introduction, in *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press:3-10.

Dillman, D.A. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley.

Dillman, D.A. 1983. Mail and Self-administered Surveys, in *Handbook of Survey Research*, editors P.H. Rossi, J.D. Wright & A.B. Anderson. [S.I.: s.n.]: 359-377.

Dillman, D.A. 1991. The Design and Administration of Mail Surveys, in *Annual Review of Sociology Volume 17*, editors W.R. Scott & J. Blake. California: Annual Reviews Inc.: 225-249.

Dillman, D.A., Sinclair, M.D. & Clark, J.R. 1993. Effects of Questionnaire Length, Respondent-Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys. *Public Opinion Quarterly*, 57:289-304.

Dohrenwend, B.S. 1970-71. An Experimental Study of Payments to Respondents. *Public Opinion Quarterly*, 34:621-624.

Donald, M.N. 1960. Implications of Nonresponse for the Interpretation of Mail Questionnaire Data. *Public Opinion Quarterly*, 24:99-114.

Drew, J.H. & Fuller, W.A. 1980. Modeling Nonresponse in Surveys with Callbacks, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:639-642.

Drew, J.H. & Fuller, W.A. 1981. Nonresponse in Complex Multiphase Surveys, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:623-628.

Dunkelberg, W.C. & Day, G.S. 1973. Nonresponse Bias and Callbacks in Sample Surveys. *Journal of Marketing Research*, 10:160-168.

Durbin, J. 1954. Non-response and Call-backs in Surveys. *Bulletin of the International Statistical Institute*, 34:72-86.

Durbin, J. 1958. Sampling Theory for Estimates Based on Fewer Individuals than the Number Selected. *Bulletin of the International Statistical Institute*, 36:113-119.

Durbin, J. 1969. Inferential Aspects of the Randomness of Sample Size in Survey Sampling, in *New Developments in Survey Sampling*, editors N.L. Johnson & H. Smith. New York: Wiley.

Durbin, J. & Stuart, A. 1954. Callbacks and Clustering in Sample Surveys: An Experimental Study. *Journal of the Royal Statistical Society*, Series A, 117:387-428.

Edgett, G.L. 1956. Multiple Regression with Missing Observations Among the Variables. *Journal of the American Statistical Association*, 51:122-131.

El-Badry, M.A. 1965. A Sampling Procedure for Mailed Questionnaires. *Journal of the American Statistical Association*, 15:209-227.

Ellis, R.A., Endo, C.M. & Armer, J.M. 1970. The Use of Potential Nonrespondents for Studying Nonresponse Bias. *Pacific Sociological Review*, 13:103-109.

Eltinge, J.L. 1992. Conditions for Approximation of the Bias and Mean Square Error of a Sample Mean Under Nonresponse. *Statistics and Probability Letters*, 15(4):267-276.

Ericson, W.A. 1967. Optimal Sample Design with Nonresponse. *Journal of the American Statistical Association*, 62:63-78.

Ernst, L.R. 1978. Weighting to Adjust for Partial Non-response, in *American Statistical Association Proceedings of the Section on Survey Research Methods.* Washington, D.C.: American Statistical Association:468-473.

Fellegi, I.P. 1963. Sampling with Varying Probabilities Without Replacement. *Journal of the American Statistical Association*, 58:183-201.

Fellegi, I.P. & Holt, D. 1976. A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71:17-35.

Ferber, R. 1966. Item Nonresponse in a Consumer Survey. *Public Opinion Quarterly*, 30:399-415.

Ferber, R. & Sudman, S. 1974. Effects of Compensation in Consumer Expenditure Studies. *Annals of Economic and Social Measurement*, 3:319-331.

Ferrari, P.W. & Bailey, L. 1981. Preliminary Results of 1980 Decennial Census Telephone Followup Nonresponse Experiment, in *American Statistical Association Proceedings of the Section of Survey Research Methods.* Washington D.C.: American Statistical Association:264-269.

Fillion, F.L. 1975. Estimating Bias Due to Nonresponse in Mail Surveys. *Public Opinion Quarterly*, 39:482-492.

Fillion, F.L. 1976. Exploring and Correcting for Nonresponse Bias Using Follow-ups of Nonrespondents. *Pacific Sociological Review*, 19:401-408.

Folsom, R. 1981. The Equivalence of Generalized Double Sampling Regression Estimators, Weight Adjustments, and Randomized Hot Deck Imputations, in *American Statistical Association Proceedings of the Section on Survey Research Methods.* Washington, D.C.: American Statistical Association:400-405.

Ford, B.L. 1976. Missing Data Procedures: A Comparative Study, in *American Statistical Association Proceedings of the Social Statistics Section: Part I.* Washington, D.C.: American Statistical Association:324-329.

Ford, B.L. 1983. An Overview of Hot-Deck Procedures, in *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press:185-207.

Ford, B.L., Kleweno, D.G. & Tortora, R.D. 1980. The Effects of Procedures which Impute for Missing Items: A Simulation Study Using an Agricultural Survey, in *American Statistical Association Proceedings of the Section on Survey Research*. Washington, D.C.: American Statistical Association: 251-256.

Freund, J.E. & Walpole, R.E. 1987. *Mathematical Statistics*. 4th edition. Englewood Cliffs: Prentice Hall.

Fuller, C.H. 1974. Weighting to Adjust for Survey Nonresponse. *Public Opinion Quarterly*, 38:239-246.

Gannon, M.J., Nothern, J.C. & Carroll, S.J. 1971. Characteristics of Nonrespondents Among Workers. *Journal of Applied Psychology*, 55:586-588.

Godambe, V.P. 1955. A Unified Theory of Sampling from Finite Populations. *Journal of the Royal Statistical Society*, Series B, 17:269-278.

Goudy, W.J. 1976. Nonresponse Effects on Relationships Between Variables. *Public Opinion Quarterly*, 40:360-369.

Goyder, J. 1987. *The Silent Minority*. Cambridge: Polity Press.

Greenlees, J.S., Reece, W.S. & Zieschang, K.D. 1982. Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed. *Journal of the American Statistical Association*, 77:251-261.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Gullahorn, J.E. & Gullahorn, J.T. 1963. An Investigation of the Effects of Three Factors on Response to Mail Questionnaires. *Public Opinion Quarterly*, 27:294-296.

Gunn, W.J. & Rhodes, I.N. 1981. Physician Response Rates to a Telephone Survey: Effects of Monetary Incentive Level. *Public Opinion Quarterly*, 45:109-115.

Hansen, M.H. & Hurwitz, W.N. 1946. The Problem of Non-response in Sample Surveys. *Journal of the American Statistical Association*, 41:517-529.

Hansen, M.H., Hurwitz, W.N. & Madow, W.G. 1953. *Sample Survey Methods and Theory: Volume I, Methods and Applications*. New York: Wiley.

Hansen, M.H., Madow, W.G. & Tepping, B.J. 1983. An Evaluation of Model-dependent and Probability-sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*, 78(384):776-793.

Hartley, H.O. & Hocking, R.R. 1971. The Analysis of Incomplete Data. *Biometrics*, 27:783-824.

Hawkins, D.F. 1975. Estimation of Nonresponse Bias. *Sociological Methods and Research*, 3:461-488.

Hendricks, W.A. 1949. Adjustment for Bias Caused by Non-response in Mailed Surveys. *Agricultural Economics Research*, 1:52-56.

Herzog, T.N. & Rubin, D.B. 1983. Using Multiple Imputations to Handle Nonresponse in Sample Surveys, in *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press: 209-245.

Hilgard, E.R. & Payne, S.L. 1944. Those Not At Home: Riddle for Pollsters. *Public Opinion Quarterly*, 8:254-261.

Hochstim, J.R. & Athanasopoulos, D.A. 1970. Personal Follow-up in a Mail Survey: Its Contribution and its Cost. *Public Opinion Quarterly*, 34:69-81.

Holt, D. & Smith, T.M.F. 1979. Post Stratification. *Journal of the Royal Statistical Society*, Series A, 142:33-46.

Jackson, J.E. & Rao, P.S.R.S. 1983. Estimation Procedures in the Presence of Nonresponse, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:273-276.

Jagers, P. 1986. Post-Stratification against Bias in Sampling. *International Statistical Review*, 54:159-167.

James, J.M. & Bolstein, R. 1990. The Effects of Monetary Incentives and Follow-up Mailings on the Response Rate and Response Quality in Mail Surveys. *Public Opinion Quarterly*, 54:356-361.

James, J.M. & Bolstein, R. 1991. Large Monetary Incentives and Their Effect on Mail Survey Response Rates. *Public Opinion Quarterly*, 56:442-453.

Jobber, D. & Sanderson, S. 1981. The Effects of a Prior Letter and Coloured Questionnaire Paper on Mail Survey Response Rates. *Journal of the Market Research Society*, 25:339-349.

Kalsbeek, W.D. 1980. A Conceptual Review of Survey Error Due to Nonresponse, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:131-136.

Kalton, G. 1983a. *Compensating for Missing Survey Data: Research Report Series, Institute for Social Research*. Ann Arbor: University of Michigan.

Kalton, G. 1983b. Models in the Practice of Survey Sampling. *International Statistical Review*, 51:175-188.

Kalton, G. & Kasprzyk, D. 1982. Imputing for Missing Survey Responses, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:22-31.

Kalton, G. & Kasprzyk, D. 1986. The Treatment of Missing Survey Data. *Survey Methodology*, 12:1-16.

Kalton, G. & Kish, L. 1981. Two Efficient Random Imputation Procedures, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:146-153.

Kalton, G. & Kish, L. 1984. Some Efficient Random Imputation Methods. *Communications in Statistics - Theory and Methods*, 13(16):1919-1939.

Kanuk, L. & Berenson, C. 1975. Mail Surveys and Response Rates: A Literature Review. *Journal of Marketing Research*, 12:440-453.

Kasprzyk, D. & Lininger, C.A. 1981. The 1982 Survey of Income and Program Participation, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:122-127.

Kass, G.V. 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29:199-217.

Kerachsky, S.H. & Mallar, C.D. 1981. The Effects of Monetary Payments on Survey Responses: Experimental Evidence from a Longitudinal Study of Economically Disadvantaged Youths, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:258-263.

Kernan, J.B. 1971. Are "Bulk-Rate Occupants" Really Unresponsive? *Public Opinion Quarterly*, 35:420-422.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Kish, L. & Hess, I. 1959. A "Replacement" Procedure for Reducing the Bias of Nonresponse. *American Statistician*, 13(4):17-19.

Kivlin, J.E. 1965. Contributions to the Study of Mail-back Bias. *Rural Sociology*, 30:322-326.

Konijn, H.S. 1981. Biases, Variances and Covariances of Raking Ratio Estimators for Marginal and Cell Totals and Average of Observed Characteristics. *Metrika*, 28:109-121.

Koo, H.P., Ridley, J.C., Piserchia, P.V., Dawson, D.A., Bachrach, C.A., Holt, M.I. & Horvitz, D.G. 1976. An Experiment on Improving Response Rates and Reducing Call Backs in Household Surveys, in *American Statistical Association Proceedings of the Social Statistics Section*. Washington, D.C.: American Statistical Association:491-494.

Koopman, R.F. 1976. Fast Regression Estimates of Missing Data. *Psychometrika*, 41:277.

Krewski, D., Platek, R. & Rao, J.N.K. (eds). 1981. *Current Topics in Survey Sampling*. New York: Academic Press.

Kruskall, W. & Mosteller, F. 1980. Representative Sampling, IV: the History of the Concept in Statistics, 1895-1939. *International Statistical Review*, 48:169-195.

Kulka, R.A., Shirey, S.J., Moore, R.P. & Woodbury, N. 1981. A Factorial Experiment on the Responses of Professional Nurses to a National Mail Survey, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association: 270-278.

Kviz, F.J. 1977. Toward a Standard Definition of Response Rate. *Public Opinion Quarterly*, 41:265-267.

Laake, P. 1986. Optimal Estimates and Optimal Predictors of Finite Population Characteristics in the Presence of Nonresponse. *Metrika*, 2:69-77.

Lanke, J. 1983. Hot Deck Imputation Techniques that Permit Standard Methods for Assessing Precision of Estimates. *Statistical Review*, 21:105-110.

Lehtonen, R. & Pahkinen, E.J. 1995. *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley.

Lessler, J.T. & Kalsbeek, W.D. 1992. *Nonsampling Error in Surveys*. New York: Wiley.

Lievesley, D. 1983. Reducing Unit Non-response in Interview Surveys, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association: 295-299.

Little, R.J.A. 1982. Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association*, 77:237-250.

Little, R.J.A. 1986. Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review*, 54:139-157.

Little, R.J.A. 1988. Missing-data Adjustments in Large Surveys. *Journal of Business and Economic Statistics*, 6:287-301.

Little, R.J.A. & Rubin, D.B. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Madow, W.G., Nisselson, H. & Olkin, I. (eds.). 1983. *Incomplete Data in Sample Surveys Volume 1: Report and Case Studies*. New York: Academic Press.

Madow, W.G. & Olkin, I. (eds.). 1983. *Incomplete Data in Sample Surveys Volume 3: Proceedings of the Symposium*. New York: Academic Press.

Madow, W.G., Olkin, I. & Rubin, D.B. (eds.). 1983. *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*. New York: Academic Press.

Mandell, L. 1974. When to Weight: Determining Nonresponse Bias in Survey Data. *Public Opinion Quarterly*, 38:247-252.

Massey, J.T., Barker, P.R. & Hsiung, S. 1981. An Investigation of Response in a Telephone Survey, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:426-431.

Mayer, C.S. & Pratt, R.W. 1966-67. A Note on Nonresponse in a Mail Survey. *Public Opinion Quarterly*, 30:637-646.

Messmer, D.J. & Seymour, D.T. 1982. The Effects of Branching on Item Nonresponse. *Public Opinion Quarterly*, 46:270-277.

Morgan, J.N. & Sonquist, J.A. 1963. Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58:415-435.

Moser, C.A. & Kalton, G. 1971. *Survey Methods in Social Investigation*. 2nd edition. London: Heinemann Educational Books.

Murthy, M.N. 1983. A Framework for Studying Incomplete Data with a Reference to the Experience in Some Countries of Asia and the Pacific, in *Incomplete Data in Sample Surveys Volume 3: Proceedings of the Symposium*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press: 7-24.

Namboodiri, N.K. (ed.). 1978. *Survey Sampling and Measurement.* New York: Academic Press.

Neyman, J. 1934. On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97:558-625.

Neyman, J. 1938. Contribution to the Theory of Sampling Human Populations. *Journal of the American Statistical Association*, 33:101-116.

Nisselson, H. 1983. Overview, in *Incomplete Data in Sample Surveys Volume 1: Report and Case Studies*, editors W.G. Madow, H. Nisselson & I. Olkin. New York: Academic Press: 107-120.

Ognibene, P. 1971. Correcting Nonresponse Bias in Mail Questionnaires. *Journal of Marketing Research*, 8:233-235.

Oh, H.L. & Scheuren, F.J. 1978a. Multivariate Raking Ratio Estimation in the 1973 Exact Match Study, *in American Statistical Association Proceedings of the Section on Survey Research Methods.* Washington, D.C.: American Statistical Association:716-722.

Oh, H.L. & Scheuren, F.J. 1978b. Some Unresolved Issues in Raking Ratio Estimation, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:723-728.

Oh, H.L. & Scheuren, F.J. 1980. Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data, in *American Statistical Association Proceedings of the Section on Survey Research Methods.* Washington, D.C.: American Statistical Association:416-420.

Oh, H.L. & Scheuren, F.J. 1983. Weighting Adjustment for Unit Nonresponse, in *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press:143-184.

O'Neil, M. 1979. Estimating the Nonresponse Bias Due to Refusals in Telephone Surveys. *Public Opinion Quarterly*, 218-232.

Pavalko, R.M. & Lutterman, K.G. 1973. Characteristics of Willing and Reluctant Respondents. *Pacific Sociological Review*, 16:463-476.

Paxson, M.C., Dillman, D.A. & Tarnai, J. 1995. Improving Response to Business Mail Surveys, in *Business Survey Methods*, editors B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott. New York:Wiley: 303-316.

Platek, R. 1977. Some Factors Affecting Non-response. *Survey Methodology*, 3(2):191-214.

Platek, R. 1980. Causes of Incomplete Data, Adjustments and Effects. *Survey Methodology*, 6:93-132.

Platek, R. & Gray, G.B. 1978. Non-response and Imputation. *Survey Methodology*, 4:144-177.

Platek, R. & Gray, G.B. 1979. Methodology and Application of Adjustments for Non-response. *Bulletin of the International Statistical Institute*, 48:533-557.

Platek, R. & Gray, G.B. 1986. On the Definitions of Response Rates. *Survey Methodology*, 12(1):17-27.

Politz, A. & Simmons, W. 1949. An Attempt to get the "Not At Homes" into the Sample without Callbacks. *Journal of the American Statistical Association*, 44:9-31.

Politz, A. & Simmons, W. 1950. Note on "An Attempt to get the Not At Homes into the Sample without Callbacks". *Journal of the American Statistical Association*, 45:136-137.

Raj, D. 1956. Some Estimators in Sampling with Varying Probabilities Without Replacement. *Journal of the American Statistical Association*, 51:269-284.

Raj, D. 1968. *Sampling Theory*. New York: McGraw-Hill.

Rao, J.N.K. 1973. On Double Sampling for Stratification and Analytical Surveys. *Biometrika*, 60:125-133.

Rao, P.S.R.S. 1983a. Callbacks, Follow-ups, and Repeated Telephone Calls, in *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press:33-44.

Rao, P.S.R.S. 1983b. Randomization Approach, in *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press:97-105.

Rollins, M. 1940. The Practical Use of Repeated Questionnaire Waves. *Journal of Applied Psychology*, 24:770-772.

Rosenbaum, P.R. & Rubin, D.B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70:41-55.

Rosenthal, R. & Rosnow, R.L. 1975. *The Volunteer Subject*. New York: Wiley.

Royall, R.M. 1971. Linear Regression Models in Finite Population Sampling Theory, in *Foundations of Statistical Inference*, editors V.P. Godambe & D.A. Sprott. Toronto: Holt, Rinehart & Winston.

Royall, R.M. & Eberhardt, K.R. 1975. Variance Estimates for the Ratio Estimator. *Sankhyā : The Indian Journal of Statistics*, Series C, 37:43-52.

Rubin, D.B. 1976. Inference and Missing Data. *Biometrika*, 63:581-592.

Rubin, D.B. 1977. Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, 72:538-543.

Rubin, D.B. 1983. Conceptual Issues in the Presence of Nonresponse, in *Incomplete Data in Sample Surveys Volume 2: Theory and Bibliographies*, editors W.G. Madow, I. Olkin & D.B. Rubin. New York: Academic Press: 123-142.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Sande, I.G. 1982. Imputation in Surveys: Coping with Reality. *The American Statistician*, 36:145-152.

Sande, I.G. 1983. Hot-deck Imputation Procedures, in *Incomplete Data in Sample Surveys Volume 3: Proceedings of the Symposium*, editors W.G. Madow & I. Olkin. New York: Academic Press: 339-349.

Santos, R. 1981. Effects of Imputation on Regression Coefficients, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:140-145.

Särndal, C.E. 1978. Design-based and Model-based Inference in Survey Sampling. *Scandinavian Journal of Statistics*, 5:27-52.

Särndal, C.E. 1980. On π Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling. *Biometrika*, 67:639-650.

Särndal, C.E. 1982. Implications of Survey Design for Generalized Regression Estimation of Linear Functions. *Journal of Statistical Planning and Inference*, 7:155-170.

Särndal, C.E. 1984. Design-consistent Versus Model-dependent Estimation for Small Domains. *Journal of the American Statistical Association*, 79:624-631.

Särndal, C.E. & Swensson, B. 1985. Incorporating Nonresponse Modelling in a General Randomization Theory Approach. *Bulletin of the International Statistical Institute*, 51(3):15.2-1 -- 15.2-15.

Särndal, C.E. & Swensson, B. 1987. A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. *International Statistical Review*, 55:279-294.

Särndal, C.E., Swensson, B. & Wretman, J. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Scott, C. 1961. Research on Mail Surveys. *Journal of the Royal Statistical Society*, Series A, 124:143-205.

Shih, W.P. 1983. Nonresponses to Income Questions in Telephone Surveys, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:283-288.

Sirken, M. 1975. Evaluation and Critique of Household Sample Surveys of Substance Use, in *Alcohol and Other Drug Use and Abuse in the State of Michigan, Final Report*. Prepared by the Office of Substance Abuse Services, Michigan Department of Public Health.

Smith, T.M.F. 1976. The Foundations of Survey Sampling: a Review. *Journal of the Royal Statistical Society*, Series A, 139:183-195.

Srinath, K.P. 1971. Multiphase Sampling in Nonresponse Problems. *Journal of the American Statistical Association*, 16:583-586.

Stanton, F. [S.a.]. Notes on the Validity of Mail Questionnaire Returns. *Journal of Applied Psychology*, 23:104-115.

Stoker, D.J. 1984. *Sampling in Practice: Occasional Paper no. 11*. Pretoria: Human Sciences Research Council.

Stoker, D.J. 1985. *Die Hantering van Item Nie-respons: Seminaarlesing 85-01*. Pretoria: Universiteit van Pretoria.

Stoker, D.J. 1986. Evaluasie van die Steekproefrealisasie, in *Opnamemetodes en - praktyk*, redakteur P.C. Smit. Pretoria: Raad vir Geesteswetenskaplike Navorsing:147-165.

Stoker, D.J. 1988. *Die Analise van Komplekse Steekproefdata: Verslag WS-40.* Pretoria: Raad vir Geesteswetenskaplike Navorsing.

Stopher, P. & Sheskin, I. 1981. A Method for Determining and Reducing Nonresponse Bias, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:252-257.

Suchman, E.A. 1962. An Analysis of "Bias" in Survey Research. *Public Opinion Quarterly*, 26:102-111.

Sudman, S. 1976. *Applied Sampling*. New York: Academic Press.

Sukhatme, P.V. 1954. *Sampling Theory of Surveys with Applications*. New Delhi: Indian Society of Agricultural Statistics.

Swensson, B. 1982. A Survey of Nonresponse Terms. *Statistical Journal of the United Nations ECE*, 1:241-251.

Thomsen, I. 1973. A Note on the Efficiency of Weighting Subclass Means to Reduce the Effects of Non-response when Analyzing Survey Data. *Statistisk Tidskrift*, 4:278-285.

Thomsen, I. 1978. A Second Note on the Efficiency of Weighting Subclass Means to Reduce the Effects of Non-response when Analyzing Survey Data. *Statistisk Tidskrift*, 16:191-196.

Thomsen, I. & Siring, E. 1983. On the Causes and Effects of Nonresponse: Norwegian Experiences, in *Incomplete Data in Sample Surveys Volume 3: Proceedings of the Symposium*, editors W.G. Madow & I. Olkin. New York: Academic Press: 25-59.

Tupek, A.R. & Richardson, W.J. 1978. Use of Ratio Estimates to Compensate for Nonresponse Bias in Certain Economic Surveys, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:197-202.

Unisa *see* University of South Africa

University of South Africa. Department of Statistics. 1992. *Sample Surveys: Only Study Guide for STA206-R.* Revised Edition. Pretoria.

Vacek, P.M. & Takamaru, A. 1980. An Examination of the Nearest Neighbour Rule for Imputing Missing Values, in *American Statistical Association Proceedings of the Statistical Computing Section.* Washington, D.C.: American Statistical Association:326-331.

Waksberg, J. 1978. Sampling Methods for Random Digit Dialling. *Journal of the American Statistical Association*, 73(361):40-46.

Ward, J.C., Russick, B. & Rudelius, W. 1985. A Test of Reducing Callbacks and Not-at-home Bias in Personal Interviews by Weighting At-home Respondents. *Journal of Marketing Research*, 22:66-73.

Warner, S.L. 1965. Randomized Response: a Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 57:622-627.

Warwick, D.P. & Lininger, C.A. 1975. *The Sample Survey: Theory and Practice*. New York: McGraw-Hill.

White, A.A. 1983. Response Rate Calculation in RDD Telephone Health Surveys: Current Practices, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association: 277-282.

Wilcox, J.B. 1977. The Interaction of Refusal and Not-at-home Sources of Nonresponse Bias. *Journal of Marketing Research*, 14:592-597.

Williams, R.L. & Folsom, R.E. 1981. Weighted Hot-deck Imputation of Medical Expenditures Based on a Record Check Subsample, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:406-411.

Williams, S.R. & Folsom, R.E. 1977. *Bias Resulting from School Nonresponse: Methodology and Findings*. Prepared by the Research Triangle Institute for the National Center for Educational Statistics.

Williams, W.H. 1968. The Systematic Bias Effects of Incomplete Responses, in *American Statistical Association Proceedings of the Social Statistics Section*. Washington, D.C.: American Statistical Association:308-312.

Wiseman, F. & McDonald, P. 1979. Noncontact and Refusal Rates in Consumer Telephone Surveys. *Journal of Marketing Research*, 16:478-484.

Woltman, H.F., Miskura, S.M., Thompson, J.H. & Bounpane, P.A. 1981. 1980 Census Weighting and Variance Estimation Studies, Design and Methodology, in *American Statistical Association Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association:164-169.

Wright, R.L. 1983. Finite Population Sampling with Multivariate Auxiliary Information. *Journal of the American Statistical Association*, 78:879-884.

Yates, F. 1981. *Sampling Methods for Censuses and Surveys*. New York: Oxford University Press.