University of Southampton, 1982

FACULTY OF SOCIAL SCIENCES

DEPARTMENT OF SOCIAL STATISTICS

MULTIVARIATE ANALYSIS OF SAMPLE SURVEY DATA

by

Christopher John Skinner

Thesis submitted for the degree of Doctor of Philosophy

To Maggie, Thomas and Samuel.

CONTENTS

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCES

SOCIAL STATISTICS

Doctor of Philosophy

MULTIVARIATE ANALYSIS OF SAMPLE SURVEY DATA

by Christopher John Skinner

Multivariate methods are used widely with sample survey data, yet the assumption of independently and identically distributed observations underlying many of these methods may be invalid for surveys of complex design. This thesis attempts to outline a formal statistical approach to this problem.

A distinction is drawn between a disaggregated approach, where the aim is to model the data in relation to the structure of the population used in the sample design, and an aggregate approach where the target of inference is a population characteristic. Only the latter approach is considered. Most attention is given to the choice and properties of point estimators of a covariance matrix. In addition the estimation of correlation coefficients, regression coefficients, principal components and parameters in factor analysis is considered.

Inference is mainly based on stochastic superpopulation models rather than on the classical randomisation distribution induced by a probability sampling design. The thesis divided into two parts. In the first part, a very general sample selection scheme depending on a set of design variables is combined with a rather restrictive classical superpopulation model in which units are independent with values distributed multivariate-normally. In the second part, a conventional two-stage sampling scheme is combined with a general superpopulation model for a clustered population.

CHAPTER ONE - INTRODUCTION

1.1 The Use of Multivariate Analysis in Social Survey Research

Multivariate methods are now widely used in the social sciences for the analysis of social survey data. Recent books on the use of such methods for the analysis of surveys are Ferber (1980) and O'Muircheartaigh and Payne (1977a, b). The contents of a haphazard sample of four recent social science journals are analysed in Table 1.1. In three of these journals over 60% of the papers contain some multivariate analysis (usually correlation analysis, regression analysis or factor analysis) and of these the majority are based on social survey data.

Table 1.1. Contents of Four Recent Social Science Journals.

Paper includes	Human Relations 33 1980	J. of Marriage & The Family 41 1979	Social Forces 58 1979/80	Sociology & Social Research 64 1979/80
no empirical work	15	13	13	4
empirical work[1] but no multivariate analysis	14	16	10	5
multivariate analysis based on	23	45	45	18
(a) social survey data	(16)	(38)	(28)	(11)
(b) not social survey data[2]	(7)	(7)	(17)	(7)
Total number of papers	52	74	68	27

Notes: [1] Includes comparisons of subclass means, simple cross-tabulations, standardised rates.

[2] Includes studies where the units of analysis are countries, U.S. states etc. and small experimental studies.

One might offer two related reasons why multivariate methods are important for the analysis of social survey data.

(1) Measurement: Many 'concepts' in the social sciences do not possess unique operational definitions and instead several indicators are often measured which are subsequently analysed simultaneously using, for example, factor analysis.

(2) Explanation: Social research is frequently concerned with analysing relationships between variables using multivariate methods, such as regression analysis. These methods are particularly important for social surveys because they can provide 'statistical control' in place of experimental control.

In this thesis we shall view multivariate analysis as a branch of statistical inference; observations are assumed to be generated by a stochastic *model*. The mechanism by which a given model generates the observations may have a theoretical justification or it may just be a mathematical simplification (Bartholomew, 1973 pp.1-9). The usual aim of multivariate analysis is to make inference about certain parameters of the model.

Conventional models in multivariate analysis (e.g. Morrison, 1976) are either *unstructured*, where observations $y_1 \ldots y_n$ are realisations of independently and identically distributed (IID) random variables, $Y_1 \ldots Y_n$, with probability densities $f(y_i|\Theta)$ $(i=1 \ldots n)$, or *structured*, where $y_1 \ldots y_n$ are realisations of random variables $Y_1 \ldots Y_n$ which are independent conditionalon known values $x_1 \ldots x_n$ with densities $f(y_i|x_i, \Theta)$ $(i=1 \ldots n)$. In each case Θ is the object of interest.

Models for observations from sample surveys will be discussed in Section 1.2 but are broadly of two types:

(1) Observations are generated by what Cassel et al (1977 Ch.2) call the Fixed Population Model in which the only stochastic element enters via a probability sampling design and the parameters of the model are the finite population values.

(2) The finite population values are themselves generated by a stochastic model, the *superpopulation model*. In this case the model generating the sample obsevations might be taken as either (a) the combination of the superpopulation model

and the sampling design or (b) if the design is non-informative, just the superpopulation model restricted to the sampled units (e.g. Royall, 1971).

For the Fixed Population Model (1), the sample observations will only obey the unstructured model if the design is srswr and only obey the structured model if the design is stratified srswr and stratum membership is denoted by x_i. Similarly for model (2)(a) the sample observations will only obey the unstructured model if the superpopulation model is unstructured and the design is non-informative and only obey the structured model if the superpopulation model is structured with respect to x_i and the design is non-informative given the x_i. In other cases the conventional models of multivariate analysis will not apply. Even in these cases, it is by no means obvious that conventional methods are applicable because the target for inference may no longer be the 'conventional parameter of interest' for the sample model. We discuss possible targets for inference in Section 1.3. but we do now give two examples.

(i) Suppose in the superpopulation model that $Y_1 \ldots Y_n$ are IID, $N(\mu, \sigma^2)$ where μ is the target for inference. Suppose the (informative) sampling design selects only those units in the population for which $y_i < K$. Then the sample model is IID (truncated normal) with a mean which is no longer the target of inference.

(ii) In some circumstances we may be interested in the specific finite population (see 1.3.1) rather than the superpopulation model. In this case under approach (2) above, conventional parametric inference would be inappropriate since the object of inference would be realisation of a random variable rather than a parameter.

Formally, there seem to be three possible approaches to the multivariate analysis of sample survey data.

(a) We could treat the multivariate anlysis as an exercise in data analysis/descriptive statistics eschewing sample/population distinctions. This is an approach that has attracted increasing interest (e.g. Gnanadesikan, 1977) due mainly to the unrealistic

multivariate normality assumptions of classical multivariate analysis.

(b) We might define finite population parameters of interest e.g. correlation matrices, covariance matrices, linear regression coefficients and consider inference about these parameters under the Fixed Population Model (1) above.

(c) We might adopt the superpopulation model approach (2) above and make inference either about the finite population 'parameters' as in (b) or about the parameters in the super-population model.

The choice of approach depends to a large extent on the context of the analysis. When analysing a pilot survey or when 'searching for structure' at an early stage of analysis approach (a) might be sensible. We are, however, specifically interested in the sampling mechanism and so shall reject approach (a).

We now argue why we prefer approach (c) to approach (b). Firstly, there are compelling reasons (e.g. Royall, 1971), which we shall not pursue, why approach (c) is more desirable than (b) even for the classical survey sampling problem of estimating finite population means and totals. More importantly, we consider that there is usually a qualitative difference between multivariate analysis and this classical problem. In studying relationships between variables and the effects of measurement error etc. it is almost *necessary* to have a model. Even a simple statistic such as a product-moment correlation coefficient lacks meaning if there is no underlying linear relationship. Hansen et al (1978) as well as Särndal and Kempthorne, in discussion of their paper, refer to the necessity of a model in causal analysis. Särndal adds that 'it seems clear to me that the model-based framework, being of wider scope, will prove superior in the development of this area' (data analysis for sample surveys). Several discussants of Kish and Frankel (1974) also question the use of approach (b) for regression analysis on the same grounds. Finally, although the use of approach (b) might be reasonable for certain descriptive multivariate analysis, our consistent use of approach (c) will provide a more uniform theoretical perspective.

In this thesis we shall be addressing two broad questions (discussed in a more restricted form in Section 1.2.2):

(A) To what extent are conventional multivariate methods, specifically those based on unstructured IID models, applicable to sample survey data?

(B) What alternative methods might be adopted which are more appropriate for sample survey data?

In particular we shall be concerned with (i) the implications of non-independence in the superpopulation model due to clustering and (ii) the effect of selection with respect to variables correlated with variables of interest in an IID model.

We might compare questions (A) and (B) with the more 'traditional' questions posed, for example, by Kish and Frankel (1974). They note that most statistical methods are based on the assumption of srs and ask what is the impact of complex survey designs. We suggest that often their questions are included in ours since complex designs will only usually be adopted if an IID model is inappropriate. Our questions are, however, more general because as Kempthorne (1978) notes, for example: 'That one should pay attention to clustering or covariance in attempted causal modelling *even if* one has a srs, seems obvious'.

In the classical problem of estimating a finite population mean question (A), i.e. what are the properties of the (unweighted) sample mean as an estimator of the finite population mean, is fairly trivial and much of survey sampling theory (e.g. Cochran, 1977) is devoted to question (B), i.e. how should we best estimate the finite population mean for given designs and/or models. In this thesis we shall give considerable attention to question (A). Three reasons for this are:

(i) Question (A) is not so trivial for multivariate analysis.

(ii) Whereas most practicising survey samplers take account of complex survey designs when estimating population means and totals, many social scientists still use (and are likely to continue to use) IID based methods. This is sometimes for practical simplicity or due to the availability of computer packages and is sometimes forced on secondary users

of survey data who do not have access to the design information, perhaps for confidentiality reasons (see also discussion by Rao and Scott, 1981).

(iii) There are conjectures that multivariate methods are more robust to departures from the IID assumption. For example, Morgan and Sonquist (1963) wrote that 'there is some reason to believe that the clustering and stratification of the sample becomes less and less important the more complex and more multivariate the analysis being undertaken'.

We end this opening section with Smith's (1976) indication of the importance of this subject:

'The vast majority of surveys are multivariate and multipurpose. The design and analysis of multivariate surveys must be one of the next major areas for research and if theoretical statisticians fail to rise to the challenge the rift between them and practical statisticians will grow wider.'

1.2 Formal Framework for Sample Surveys

The foundations of sample survey theory have been extensively investigated (reviews are given by Cassel et al., 1977, and Smith, 1976) and so our discussion is brief and restricted to selected topics.

1.2.1 Sampling Designs

We consider a finite *population* of N identifiable units denoted by U = {1...N}. A *sample* is defined as a subset of U (hence our definition ignores the order of selection or multiplicity of units). Let $\mathcal{S}$ be the set of all subsets of U. A *sampling design*, p(s), is a real-valued function on $\mathcal{S}$ such that:

$$p(s) \geq 0 \quad \text{for all } s \in \mathcal{S}$$

$$\sum_{s \in \mathcal{S}} p(s) = 1$$

The design defines a probability distribution for a random variable S taking values $s \in \mathcal{S}$; P(S=s) = p(s).

We suppose that, associated with the i^{th} unit of U, there is a pair of vectors (y_i, x_i) of dimensions (p, q) (i=1...N). The y variables are *variables of interest (inference variables*, Smith, 1978) and are observed for members of the sample but are unobserved for other units in U. The x variables are *auxiliary variables (design variables*, Smith, 1978). Let $\underline{x} = (x_1'...x_N')'$ and $\underline{y} = (y_1'...y_N')'$. We assume that p(s) is a (deterministic) function of $\underline{x}$ and does not depend on $\underline{y}$ and write $p(s) = p(s|\underline{x})$. The identification of the units may be used in the design by letting the first component of x_i be the label, i (i=1...N). We distinguish between two cases:

(1) Known selection scheme: $\underline{x}$, N and $p(s|\underline{x})$ are all known before the sample is selected. In this case the design is said to be *non-informative* given $\underline{x}$ and N.

(2) Unknown selection scheme: Not all of $\underline{x}$, N and $p(s|\underline{x})$ are known before the sample is selected.

Examples of Known Selection Schemes

1. All the usual probability sampling designs discussed in standard sampling textbooks (e.g. Cochran, 1977) have known selection schemes. We give two examples:

(a) Example 1.1 - Stratified random sampling without replacement

Let $\underline{e}_1$ be the (H-1) vector of zeros and let $\underline{e}_h$ be the (H-1) vector with unity as its $(h-1)^{th}$ element and zeros elsewhere (h=2...H). Suppose that $\underline{x}_i$ is known and may only take one of the values $\underline{e}_1...\underline{e}_H$ (i=1...N). Then U may be partitioned into H strata $S_h = S_h(\underline{x})$ (h=1...H) such that:

$$i \in S_h \Longleftrightarrow \underline{x}_i = \underline{e}_h$$

Let N_h be the number of units in $S_h (\Sigma N_h = N)$ and let $n_1...n_H$ be given integers such that $1 \leq n_h \leq N_h$ (h=1...H). Let $\mathcal{S}_h$ be the set of all subsets of S_h of size n_h and let

$$\mathcal{S}_{ST} = \mathcal{S}_{ST}(\underline{x}) = \{s_1 U...U s_H : s_h \in \mathcal{S}_h , h=1...H\}$$

Then a stratified random sampling without replacement design is defined by

$$p(s|\underline{x}) = \prod_{h=1}^{H} \binom{N_h}{n_h}^{-1} \qquad s \in \mathcal{S}_{ST}$$

$$= 0 \qquad s \notin \mathcal{S}_{ST} \qquad (1.1)$$

(b) Example 1.2 - Two-stage Sampling

Let the $\underline{x}_i$ and S_h be defined as in (a). Let $\mathcal{S}'$ be the set of all subsets of {1...H}. Let $p'(s'|\underline{x})$ be a sampling design on $\mathcal{S}'$. Let $\mathcal{S}_h$ (now) be the set of all subsets of S_h. For each $s' \in \mathcal{S}'$ write, without loss of generality, $s' = \{1...n_{s'}\}$ and let

$$\mathcal{S}(s') = \{s_1 U ... U s_{n_{s'}} : s_h \in \mathcal{S}_h\}$$

Let $\mathcal{S}_{TS} = \bigcup_{s' \in \mathcal{S}'} \mathcal{S}(s')$

Let $p_h(s_h|\underline{x}, s')$ be a given sampling design on $\mathcal{S}(s')$. Then a two-stage sampling design is defined by

$$p(s|\underline{x}) = \left(\prod_{h \in s'} p_h(s_h|\underline{x}, s')\right) p'(s'|\underline{x}) \qquad s \in \mathcal{S}_{TS}$$

$$= 0 \qquad s \notin \mathcal{S}_{TS}$$

(Note that, as defined above, $\mathcal{S}_{TS}$ is in fact the set of all subsets of U. $\mathcal{S}_{TS}$ is usually restricted by constraints on $\mathcal{S}'$ and the $\mathcal{S}_h$).

2. Lord and Novick (1968 p.140) describe the selection of samples on the basis of test scores (or vectors of test scores) x_i. Usually s consists of those units in U such that $x_i \geq K$ where K is a specified number.

3. Royall (1970) describes an 'optimal' sampling plan in which the n units in U with largest (univariate) x_i values are selected with probability one and other units are selected with probability zero.

Examples of Unknown Selection Schemes

4. Scott (1977) describes a situation where $\underline{x}$ is unknown, e.g. secondary users may know that a stratified sample was selected but not be able to identify which strata individual members of the sample belong to.

5. The mechanism by which non-response occurs in sample surveys is generally unknown. We might suppose that the probability of response depends only on a specific set of variables x_i, but where $p(s|\underline{x})$ and $\underline{x}$ are unknown (c.f. Rubin, 1977).

1.2.2 Models and Targets for Inference

As noted in Section 1.1 we shall adopt a superpopulation model approach. Most conventional superpopulation models are cases of what we shall call a *conditional superpopulation model*. Recall $\underline{y} = (y_1' \ldots y_N')'$ and $\underline{x} = (x_1' \ldots x_N')'$. Then in a conditional superpopulation model $\underline{y}$ is assumed to be a realisation of the random vector $\underline{Y} = (Y_1' \ldots Y_N')'$ with probability density function (p.d.f) $p(\underline{Y}|\underline{x})$ conditional on $\underline{x}$. We shall, however, find that a more convenient framework is offered by an *unconditional superpopulation model* where $(\underline{y}'\ \underline{x}')'$ is a joint realisation of $(\underline{Y}'\underline{X}')'$ with joint p.d.f. $p(\underline{Y}|\underline{X})p(\underline{X})$ where $\underline{X} = (X_1' \ldots X_N')'$. This enables us also to consider the marginal distribution of $\underline{Y}$ with p.d.f. $p(\underline{Y})$. In this section we assume that $p(\underline{Y}|\underline{X})$, $p(\underline{X})$ and $p(\underline{Y})$ are members of known classes of distributions indexed by unknown (usually vector) parameters $\theta \in \Theta$, $\phi \in \Phi$ and $\psi \in \Psi$ respectively. We write $p(\underline{Y}|\underline{X},\theta)$, $p(\underline{X}|\phi)$ and $p(\underline{Y}|\psi)$. ψ is, of course, a function of θ and ϕ. For most of this thesis we shall, in fact, relax the assumption that the p.d.f.'s are members of given parametric families and only assume certain moment properties of the distributions. In Chapter 5 we shall also relax the assumption, made implicitly above, that N is fixed and allow it, as well, to be a realisation of a random variable.

The choice of targets for inference will be discussed in Section 1.3. The target will be either a function of $\underline{y}$ (a finite population 'parameter') or a function of ψ (a superpopulation parameter). In fact, for all the models that we shall consider, $Y_1 \ldots Y_N$ will possess an exchangeable distribution (unconditional on $\underline{x}$) and hence share a common marginal distribution, $p(Y|\psi_o)$ say, indexed by a parameter ψ_o, a function of ψ. The only superpopulation parameters of interest that we shall consider will be function of ψ_o.

In this thesis we shall only be concerned with point estimation (see Section 1.4). For this restricted problem we may rephrase questions (A) and (B) of Section 1.1 as:

(A) What are the properties of the standard estimators of multivariate analysis in the survey context?

(B) What alternative estimators might we adopt?

Question (A) is essentially a *robustness* question. What happens when the standard assumptions of multivariate analysis do not hold? Our use of the term robustness here is broader than, say, the restricted definition of Huber (1972), who is still concerned with IID observations, and we would include studies such as that of Praetz (1981) who considers the effect of serially correlated residuals on F-tests in multiple regression. To answer question (A) we compare the properties of 'classical' (standard) estimators under a hypothetical 'true' model, *Model I*, with the properties under a corresponding IID model, *Model II*, in which the Y_i (i=1...N) are assumed to be IID and independent of $\underline{x}$ with common distribution $p(Y|\psi_o)$ (defined above). A difference between the properties of the estimator under the two models will be interpreted as a *misspecification effect*, i.e. an effect of misspecifying the model as Model II when, in fact, the true model is Model I. This is the model-based analogy of the more usual concept of a *design effect* (e.g. Kish, 1965), the effect of using a complex sampling design instead of srs. Question (A) is not formally a problem of statistical inference. We might, for example, consider the sampling distribution of a given estimator conditional on any statistics of our choice, if this helps us to understand the properties of the estimator. We do not even need to 'know' the sample selection scheme which might, for example, be a combination of a probability design and non-response. We proceed *as if* hypothetical sampling schemes and models were correct. This kind of investigation is an example of the traditional use of super-population models (Smith, 1976).

Question (B) essentially implies an *optimality* question. What is the best estimator of a given quantity? This is a classical problem of statistical inference and involves the more recent use of super-population models (Smith, 1976). It assumes the model to be correct

and if it is not we should ideally investigate the effect of departures from the model on the optimal estimators.

1.2.3 Inference

Likelihood Approach

Without loss of generality let $s = \{1...n\}$

Let $\underline{y}_s = (y_1'...y_n')'$; $\underline{y}_{\bar{s}} = (y'_{n+1}...y'_N)'$

$$p(\underline{y}_s|\underline{x},s,\theta) = \int p(\underline{y}|\underline{x},\theta)d\underline{y}_{\bar{s}}$$

Assuming a known selection scheme the data is

$$d = (\underline{y}_s,\ s,\ \underline{x})$$

Hence the likelihood (for the unconditional superpopulation model) is

$$L(\theta,\ \phi) \propto p(\underline{y}_s|\underline{x},\ s,\ \theta)\ p(s|\underline{x})\ p(\underline{x}|\phi) \qquad (1.2)$$

$$\propto p(\underline{y}_s|\underline{x},\ s,\ \theta)\ p(\underline{x}|\phi),\ \text{since } p(s|\underline{x}) \text{ is known.}$$

Hence from the Likelihood Principle (e.g. Cox and Hinkley, 1974, p.39) inference about θ and ϕ, and hence ψ, should not depend on the sampling design, $p(s|\underline{x})$. (See Smith, 1978).

The Likelihood Approach to the prediction of $\underline{y}$ is more problematic. Hinkley (1979) (also Lauritzen, 1974) would define the predictive likelihood as

$$p(d|T) \propto p(\underline{y}_s|T,\ \underline{x},\ s)\ p(s|\underline{x})\ p(\underline{x}|T)$$

$$\propto p(\underline{y}_s|T,\ \underline{x},\ s)\ p(\underline{x}|T)$$

where $T = T(\underline{y},\ \underline{x})$ is a minimal sufficient statistic for $(\theta,\ \phi)$ were $\underline{y}$ to be observed. Royall (1976a) would define the predictive likelihood of a function, $h(\underline{y})$, of $\underline{y}$ of interest as

$$p(\underline{y}_s|h(\underline{y}),\ s,\underline{x},\theta)p(\underline{x}|h(\underline{y}),\ \phi)$$

this being the ratio of the 'posterior distribution' of $h(\underline{y})$, $p(h(\underline{y})|d)$, and the 'prior distribution' $p(h(\underline{y}))$. An advantage of Hinkley's definition is that the likelihood does not depend on the parameters $(\theta,\ \phi)$, whereas Royall's likelihood will in general. A disadvantage of Hinkley's definition is that his likelihood may often

be degenerate. In each case, however, inference about $\underline{y}$ or $h(\underline{y})$ does not depend on the sampling design.

Bayesian Approach

Let $\tau(\theta, \phi)$ be a prior distribution for (θ, ϕ). The posterior distribution of (θ, ϕ) is

$$p(\theta, \phi|d) \propto p(\underline{y}_s|\underline{x}, s, \theta)\ p(s|\underline{x})\ p(\underline{x}|\phi)\ \tau(\theta, \phi)$$

$$\propto p(\underline{y}_s|\underline{x}, s, \theta)\ p(\underline{x}|\phi).\tau(\theta, \phi)$$

The posterior distribution of a function $h(\underline{y})$ of $\underline{y}$ is

$$p(h(\underline{y})|d) \propto p(\underline{y}_s|h(\underline{y}), \underline{x}, s)\ p(s|\underline{x})\ p(h(\underline{y})|\underline{x})\ p(\underline{x})$$

$$\propto \int p(\underline{y}_s|h(\underline{y}), \underline{x}, s, \theta)\ p(h(\underline{y})|\underline{x}, \theta)\ p(\underline{x}|\phi)\ \tau(\theta, \phi)\ d\theta\ d\phi$$

Note that if θ and ϕ are prior independent so that $\tau(\theta, \phi) = \upsilon(\theta)\eta(\phi)$ then

$$p(h(\underline{y})|d) \propto \int p(\underline{y}_s|h(\underline{y}), \underline{x}, s, \theta)\ p(h(\underline{y})|\underline{x}, \theta)\upsilon(\theta)\ d\theta$$

i.e. inference from the unconditional superpopulation model is the same as from the conditional superpopulation model. As for the Likelihood Approach, inference about ψ or $\underline{y}$ does not depend on the sampling design.

Sampling Theory Approach

We might evaluate the sampling distribution of an estimator, $e(d)$, with respect to repeated realisations $(\underline{y}, \underline{x})$ from the model (ξ) distribution and/or with respect to the randomisation (p) distribution induced by repeatedly selecting samples using $p(s|\underline{x})$. Various combinations of the ξ and p-distributions have been used, e.g. Cassel et al (1976) consider minimising the ξp-MSE subject to p-unbiasedness, on the basis of essentially ad hoc grounds. Formally, however, it would only seem appropriate not to consider the joint ξp-distribution if we can 'separate' the inference procedure by margining to a sufficient statistic or conditioning on an ancillary statistic.

Definition 1.1 : Suppose a parameter λ, taking values in Λ, may be partitioned into $\lambda = (\theta, \phi)$ where θ takes values in Θ and ϕ takes value in Φ. θ and ϕ are said to be *Cartesian independent* if Λ is the Cartesian product of Θ and Φ, i.e.

$$\Lambda = \{(\theta, \phi) : \theta \in \Theta, \phi \in \Phi\}$$

Definition 1.2 : For a model indexed by $\lambda = (\theta, \phi)$, where θ and ϕ are Cartesian independent, suppose $S = (T,C)$ is a sufficient statistic for λ. If

(a) the p.d.f. of C depends on ϕ but not on θ,

(b) the conditional p.d.f. of T given C = c depends on θ but not on ϕ for all values of c, then C is called *ancillary* for θ.

Definition 1.2 is the definition of 'extended ancillarity' given by Cox and Hinkley (1974, p.35) and is the same as the definition of S-ancillarity given by Barndorff-Nielsen (1978, p.50). The *Conditionality Principle* (Cox and Hinkley, 1974 p.38) then states that if C is ancillary for θ then inference about θ should be made conditional on C taking its observed value. This principle is not entirely well-defined since C may not be unique but we shall not consider this problem. As an example, let C = (s, $\underline{x}$), T = $\underline{y}_s$ and θ and ϕ (supposed Cartesian independent) be as defined in the unconditional superpopulation model then C is ancillary for θ and so, according to the Conditionality Principle we should make inference about θ conditional on s, the actual sample obtained, and on $\underline{x}$.

Formal conditioning arguments for prediction appear only to be available in terms of sufficiency and not ancillarity although the Conditionality Principle is still appealed to (e.g. Royall and Cumberland, 1981, p.68). Lauritzen's (1974) definition of predictive sufficiency may be expressed as:

Definition 1.3. Let Y be an observed random vector with distribution indexed by θ. Let S=S(Y) be a sufficient statistic for θ. Let Z be an unobserved random vector such that the joint distribution of

(Y, Z) is also indexed by θ. Then S is said to be *predictive sufficient* for Z if Y is conditionally independent of Z given S.

We now propose a definition of predictive ancillarity in the spirit of Definition 1.2 and 1.3.

Definition 1.4. Let Y, S, Z and θ be as in Definition 1.3 (S is sufficient for θ and predictive sufficient for Z). Suppose $\theta = (\lambda, \phi)$ where λ and ϕ are Cartesian independent and suppose S = (T,C). Then C is said to be *predictive ancillary* for Z if:

(a) the p.d.f. of C depends on ϕ but not on θ,

(b) the conditional p.d.f. of (T, Z) given C=c depends on θ but not on ϕ for all values of c.

The analogy of the Conditionality Principle is then to make inference about Z conditional on C taking its observed value, if C is predictive ancillary for Z.

We are interested in making inference about either $\underline{y}$ or ψ in the unconditional superpopulation model. If we let $T = \underline{y}_s$, $Z = \underline{y}_{\bar{s}}$ and $C = (s, \underline{x})$ and suppose θ and ϕ are Cartesian independent then C is predictive ancillary for Z and we therefore make inference about Z and hence about $\underline{y}$ conditional on C. As in the Bayesian approach,where the condition that θ and ϕ are prior independent is equivalent to the condition that θ and ϕ are Cartesian independent in this case, inference from the unconditional superpopulation model would be the same as for the conditional superpopulation model and it does not depend on the sampling design, but only on the actual sample obtained.

Inference about ψ is more problematic.

Example 1.3: Consider Example 1.1 of Section 1.2.1. Suppose the Y_i are independent Bernoulli random variables given $\underline{x}$ with $P(Y_i=1 | \underline{x}_i=\underline{e}_h) = \theta_h$ so that

$$p(\underline{y}_s | s, \underline{x}, \theta) = \prod_{h=1}^{H} \theta_h^{m_h(s)} (1-\theta_h)^{n_h - m_h(s)} \qquad (1.3)$$

where $\theta = (\theta_1 \ldots \theta_H)$ and $m_h(s) = \sum_{i \in s: \underline{x}_i = \underline{e}_h} y_i$

Suppose that $p(s|\underline{x})$ is a proportionate allocation stratified srswor design and the $\underline{x}_i$ are IID with $P(\underline{x}_i = \underline{e}_h) = \phi_h$ $(\Sigma\phi_h = 1)$ and N is fixed. Then $p(s|\underline{x})$ is given in (1.1) and

$$p(\underline{x}|\phi) = \prod_{h=1}^{H} \phi_h^{N_h}, \quad \phi = (\phi_1 \ldots \phi_H) \tag{1.4}$$

Suppose the parameter of interest is $\psi = E(Y_i) = \Sigma\theta_h \phi_h$.

In the example above we should like to consider the properties of a given estimator (say the sample mean) of ψ, conditional on the actual sample, s, obtained rather than averaging its properties over all possible samples. We should also like to consider the properties of an estimator conditional on $\underline{x}$ since, for example, if H = 2 we would expect an estimate based on sample with $N_1 = N_2 = N/2$ to be 'better' than an estimate based on a sample with $N_1 = N$, $N_2 = 0$. However, $\underline{x}$ is sufficient for ϕ and so if C is a function of $\underline{x}$ then the distribution of the data, d, given C can depend only on θ. Since ψ is not a known function of θ, C cannot be ancillary for ψ by condition (b) of Definition 1.2. Hence we cannot appeal to the Conditionality Principle to make inference about ψ conditional on $\underline{x}$. Similarly, the marginal distribution of s depends on a function $a(\phi)$ of ϕ whereas the conditional distribution of $\underline{y}_s$ and $\underline{x}$ given s depends on a function $b(\theta, \phi)$ of (θ, ϕ) which is not Cartesian independent of a (ϕ). Hence again from Definition 1.2 s is not ancillary for ψ and we cannot condition on s in making inference about ψ.

We might attempt to construct some ad hoc procedures for inference about ψ which avoid the use of $p(s|\underline{x})$. A point estimator of ψ may be obtained by setting $\hat{\psi} = \psi(\hat{\theta}, \hat{\phi})$ where $\hat{\theta}$ is derived from the conditional distribution, $p(\underline{y}_s|\underline{x},\theta)$ and $\hat{\phi}$ from $p(\underline{x}|\phi)$. In our example we might take $\hat{\theta}_h = m_h(s)/n_h$, $\hat{\phi}_h = N_h/N$ and hence set $\hat{\psi} = \Sigma\, \hat{\theta}_h\, \hat{\phi}_h$. We cannot construct a confidence interval for ψ of known confidence level where the confidence measure is conditional on s. However we could for example obtain a confidence interval for $\psi^* = \Sigma\theta_h\hat{\phi}_h$ which had known confidence level conditional on s and $\underline{x}$ and we could also obtain ε such that $P(|\psi^* - \psi| < \varepsilon)$ takes a given value without reference to $p(s|\underline{x})$.

We have argued above that, *in general*, it is impossible to appeal to the Conditionality Principle to condition on s when making inference about ψ. In some circumstance it will be possible, for example if $\theta_1 = \ldots = \theta_H = \theta$ above then $\psi = \theta$ and (s, $\underline{x}$) in ancillary for ψ. That fundamental differences exist between the problems of predicting $\bar{y}$ and estimating ψ is not unknown; for example, Royall and Herson (1973 p.881) note the differences between optimal design for these two cases. In a practical sense, however, these differences are annoying. For by letting N increase we may make finite population parameters arbitrarily close to the corresponding superpopulation parameters (see 1.3.2), yet formally we should make conditional inference about the former and (in general) unconditional inference about the latter. For this reason it may be unnecessarily formal not to adopt a conditional approach for estimating ψ (as in Holt et al, 1980). As in the example above we might always susbtitute ψ^* for ψ, a difference of no practical significance if N is large.

Proposed Approach

We shall mainly adopt a Sampling Theory approach. This choice is largely arbitrary and we do not intend here to discuss comparative inference. It does, however, have the advantage that we may make fairly weak model assumptions in terms of moments without specifying distributional forms and we may evaluate the properties of estimators in terms of the traditional survey sampler's measures of bias and MSE. It does have the disadvantage, discussed above, that conditional inference about ψ is problematic.

We note that our choice of point estimators would not vary much between approaches. We shall especially use maximum likelihood estimators which have a natural interpretation in the Likelihood Approach, are posterior modes with respect to uniform priors in the Bayesian Approach and have optimal asymptotic properties in the Sampling Theory Approach.

1.3 Object of Inference

1.3.1 Finite or Superpopulation Parameters

The task of multivariate analysis is to represent complex sets of data in a simple and 'interpretable' way. In Section 1.1 we noted that the approach of classical multivariate analysis to this task depends fundamentally on the specification of a model such that (i) the model has a simple structure and has 'interpretable' parameters and (ii) the data are consistent with the hypothesis that the data is a realisation from the model. The objects of inference are the parameters of the model.

For our problem, where a sample is selected from a finite population, it is most natural and analogous to view the superpopulation model as the data generating mechanism of interest and to view the probability sampling design and the realisation of the finite population as impositions on top of the model which do not alter our objects of interest. In this case the targets for inference will be the superpopulation parameters.

It may be argued, however, that in certain circumstances finite population 'parameters' will be of interest. In time series analysis it is assumed that a given time series is a single realisation of a stochastic process and, although the parameters of this stochastic process may be interesting *per se*, when making forecasts one is interested in future values for the given realisation rather than the future behaviour of the model. Similarly one might be interested in the actual correlation coefficient for a given finite population rather than for the hypothetical superpopulation from which the finite population is a 'sample'. Fuller (1973) notes that there is a third possibility. We may be interested in the 'parameters' of a finite population separated from the finite population studied by time or space. We might assume that both these finite populations are independent realisations of the same superpopulation model.

Inferences about finite or superpopulation parameters are sometimes

called *descriptive* or *analytical* inferences respectively (e.g. Holt and Smith, 1976). On the other hand these terms are also taken to refer to non-causal or causal analyses respectively (e.g. Rao, 1975). That these two 'definitions' are equivalent is not obvious. It may be that for causal inference only superpopulation parameters are relevant (e.g. Barnard, 1971; Kalton, 1976; Hansen et al, 1978) but superpopulation parameters must also be of interest in non-causal analyses such as factor analysis.

Overall we suspect that superpopulation parameters are of most relevance for multivariate analysis and we take this to be also the broad conclusion of Fuller's (1973) useful discussion of regression analysis. This approach, we suggest, is most likely to appeal to those users of classical multivariate methods who have limited interest in survey sampling. On the other hand the topic of this thesis falls very much within the statistical subdiscipline of 'Survey Sampling' and so it will be useful to consider the problem of estimating finite population parameters in order to provide analogues with the classical theory of estimating means and totals.

The problem of defining a natural 1-1 correspondence between superpopulation and finite population parameters for a given model is discussed in the next section. We might hope that such a correspondence would imply that the difference between the two types of parameters converged to zero as the finite population size increased and for this reason the distinction between finite and superpopulation targets of inference should have limited practical significance.

1.3.2 Aggregated or Disaggregated Parameters

Later in this section we argue that in different circumstances either disaggregated (e.g. within-stratum) parameters or aggregated parameters may be of interest.

We consider initially the problem of definition. An example is helpful.

Example 1.4

Consider the model in Example 1.3. $\underline{x}$ partition U into subgroups $S_h = S_h(\underline{x})$ (h=1...H) where

$$i \varepsilon S_h \iff x_i = e_h \quad i=1...N$$

For present purposes these subgroups may be strata or clusters. We are not here concerned with sampling. Suppose, as in Example 1.3, that the Y_i are independent Bernoulli random variables given $\underline{x}$ with $P(Y_i = 1 | X_i = e_h) = \theta_h$. Suppose also that the X_i are IID with $P(X_i = e_h) = \phi_h$. Then the Y_i (i=1...N) are unconditionally IID Bernoulli random variables with $P(Y_i = 1) = \Sigma\theta_h \phi_h = \psi$.

In this example it is natural to define the *disaggregated superpopulation parameters* as $\theta = (\theta_1 ... \theta_H)$ and the *aggregated superpopulation parameter* as ψ. Correspondingly the *disaggregated finite population parameters* may be defined as $\theta(\underline{y}) = (\bar{y}_1 ... \bar{y}_H)$ where $\bar{y}_h = \sum_{S_h} y_i/N_h$ and the *aggregated finite population parameter* as $\psi(\underline{y}) = \bar{y} = \Sigma y_i/N$. Note that as $N \to \infty$, $\theta(\underline{y})$ converges almost surely to θ and $\psi(\underline{y})$ to ψ.

This example and its corresponding definitions may be extended naturally to other situations where the Y_i are independent between subgroups, given $\underline{x}$, and for $i \varepsilon S_h$ the Y_i are IID given $\underline{x}$ with a distribution indexed by θ_h, e.g. $Y_i \sim N_p(\mu_h, \Sigma_h)$, $\theta_h = (\mu_h, \Sigma_h)$, and where $\underline{x}$ is distributed as in this example.

We now consider how these definitions may be extended to the general model $p(\underline{Y}|\underline{X}, \theta)\, p(\underline{X}|\phi)$ of Section 1.2.2. It seems natural to define the vector of *disaggregated superpopulation parameters* as θ, but it is not clear how to define corresponding finite population parameters. We would like a map from $\theta \varepsilon \Theta$ to $\theta(\underline{y}) \varepsilon R^k$ ($\theta = (\theta_1 ... \theta_k)'$). One choice would be to define $\theta(\underline{y})$ as the maximum likelihood estimator of θ were $\underline{Y} = \underline{y}$ and $\underline{X} = \underline{x}$ to be observed. This, however, confuses the definitional question with the problem of inference. In general no natural map is available. For example, if $Y_1 ... Y_N | \underline{X} \sim$ NID $(\theta, 1)$ then θ

is both the mean and median of the superpopulation. In different circumstances either the mean or the median of the finite population might be targets of inference. For our purposes we shall usually be able to define θ as a function of the moments of $p(\underline{Y}|\underline{X})$ and then define $\theta(\underline{y})$ naturally as the same function of 'corresponding' moments of $\underline{y}$, as in Example 1.4. Even if we can adequately define 'corresponding' here we shall still face problems with models which assume that the moments are structured as in factor analysis. It is difficult to conceive of a finite population analogue to a factor loading, for example, without resorting to a point estimation map. This just adds further support to the case for choosing superpopulation rather than finite population parameters (see previous section).

The problem of defining aggregate parameters is, however, to some extent reversed, it being easier in some cases to define the finite population parameters than the superpopulation parameters. If $Y_1 \ldots Y_N$ are unconditionally IID with common marginal distribution indexed by ψ, as in Example 1.4, then it seems natural to view ψ as the *aggregate superpopulation parameter*. However, as noted in Section 1.2.2, in the most general models that we shall consider $Y_1 \ldots Y_N$ are unconditionally exchangeably distributed with common marginal distribution indexed by ψ_o. It is tempting to define the aggregate superpopulation parameter as ψ_o but this raises problems as the following example shows.

Example 1.5

Let $\underline{x}$ be defined as in Example 1.4 with the same marginal distribution given by (1.4). Given $\underline{x}$ we may define $\underline{Y}_h = (Y_{i1} \ldots Y_{iN_h})'$ for $h = 1 \ldots H$ where $S_h(\underline{x}) = \{i_1 \ldots i_{N_h}\}$.

Given $\underline{x}$, suppose $\underline{Y}_1 \ldots \underline{Y}_H$ are independent and that, for $h = 1 \ldots H$, $Y_{i1} \ldots Y_{iN_h}$ are the first N_h terms of an infinite exchangeable sequence, $E_h = (Y_1 Y_2 \ldots)$ of 0-1 random variables whose distribution (given $\underline{x}$) is indexed by θ_h. This defines the distribution $p(\underline{Y}|\underline{X}, \theta)$ where $\theta = (\theta_1 \ldots \theta_H)$.

Now suppose that $P(Y_i = 1|\underline{x}_i = \underline{e}_h) = \tilde{\theta}_h$ where $\tilde{\theta}_h$ is a function of θ_h. Then as in Example 1.4

$$E(Y_i) = P(Y_i = 1) = \Sigma\tilde{\theta}_h \phi_h = \psi_o, \text{ say.}$$

Now de Finetts's theorem implies (e.g. Hall and Heyde, 1980, Theorem 7.2) that there exist random variables $Z_1 \ldots Z_H$ concentrated on $[0,1]$ such that

$$P(N_h\bar{y}_h = k|\underline{x}, Z_h) = \binom{N_h}{k} Z_h^{\,k} (1-Z_h)^{N_h-k} \quad \text{a.s.}$$

Hence if $N \to \infty$ then $N_h \to \infty$ a.s. and

$$\bar{y}_h|\underline{x}, Z_h \to Z_h \text{ a.s.}$$

Hence
$$\bar{y}|\underline{x}, Z_1 \ldots Z_H \to \Sigma N_h Z_h/N \text{ a.s.}$$

and
$$\bar{y}|Z_1 \ldots Z_h \to \Sigma\phi_h Z_h \text{ a.s.} \qquad (1.5)$$

The Z_h may be interpreted as random effects. (1.5) implies that the limit of $\bar{y}$ depends on the H realisations of $Z_1 \ldots Z_H$ and so $\bar{y}$ a.s. does not converge to $\psi_o = E(Y_i)$. If ψ_o is the superpopulation counterpart of $\bar{y}$ this contradicts our desired property that differences between corresponding finite population and superpopulation parameters should converge to zero as $N \to \infty$.

In practice such models have been proposed for clustered populations (e.g. Altham, 1976) and the problem occurs because we have held the number of clusters fixed and so forced the cluster sizes to increase. The problem is removed if we let the number of clusters increase as in the following example.

Example 1.6

Let U be partitioned into H clusters $S_1 \ldots S_H$ of sizes $N_1 \ldots N_H$ respectively. Let $N_1 \ldots N_H$ be IID realisations of a random variable ν and let $\underline{N} = (N_1 \ldots N_H)$. Define $x_1 \ldots x_N$ as in Example 1.4 such that

$$i \varepsilon S_h \iff x_i = e_h \quad i=1 \ldots N$$

Given $\underline{N}$ define

$$L(\underline{N}) = \begin{pmatrix} 1_{N_1} \otimes e_1 \\ \vdots \\ 1_{N_H} \otimes e_H \end{pmatrix}$$

i.e. $L(\underline{N})$ consists of N_1 vectors e_1 stacked on top of N_2 vectors e_2 etc. We also define N! vectors $L_\pi(\underline{N})$ which are obtained from $L(\underline{N})$ by permuting the N vectors e_h in $L(\underline{N})$. Define the conditional distribution $p(\underline{X}|\underline{N})$ by

$$p(\underline{X} = L_\pi(\underline{N})\,|\,\underline{N}) = 1/N!$$

Now as in Example 1.5 suppose $\underline{Y}_1 \ldots \underline{Y}_H$ are independent given $\underline{x}$ and $\underline{N}$ and that, for h=1...H, $Y_{i1} \ldots Y_{iNh}$ are the first N_h terms of an infinite exchangeable sequence $E_h = (Y_1 Y_2 \ldots)$ of 0-1 random variables whose distribution given $\underline{x}$ and $\underline{N}$ depends on θ_h which is a known function of N_h, $\theta_h = \theta(N_h)$ (we may restrict E_h to a finite set if ν is bounded). Then

$$P(Y_i = 1\,|\,\underline{x}_i = e_h, \underline{N}) = \tilde{\theta}(N_h), \text{ say}$$

$$\therefore \quad P(Y_i = 1\,|\,\underline{N}) = \Sigma N_h\, \tilde{\theta}(N_h)/\Sigma N_h$$

and as $H \to \infty, \psi_o = P(Y_i = 1) \to E(\nu\tilde{\theta}(\nu))/E(\nu)$ a.s.

Now $\bar{y} = \Sigma N_h\, \bar{y}_h/\Sigma N_h$

and the distribution of $\bar{y}_h$ depends only on $\theta_h = \theta(N_h)$. Hence the $N_h\, \bar{y}_h$ are IID unconditionally where

$$E(N_h \bar{y}_h) = E(N_h E(\bar{y}_h\,|\,N_h)) \qquad \text{(note } \bar{y}_h \text{ is independent of } \underline{x}\text{)}$$

$$= E(\nu\tilde{\theta}(\nu))$$

Hence as $H \to \infty$

$$\bar{y} \to E(\nu\tilde{\theta}(\nu))/E(\nu) \quad \text{a.s.}$$

Speaking roughly, the reason why $(\bar{y} - \psi_o) \to 0$ as $H \to \infty$ in Example 1.6 is that the marginal distribution of the Y_i approaches IID as $H \to \infty$ which it does not in Example 1.5, Under this kind of condition it seems reasonable to define ψ_o as the aggregate superpopulation parameter. The definition of the aggregate finite population parameter then follows as in the discussion of disaggregated parameters.

We now ignore the finite/super population distinction and

attempt to classify problems for which either disaggregated or aggregated targets may be of interest. To some extent this is a hopeless task since the object of inference depends so strongly on the substantive context. However, it is an important subject and so we do try to throw a little light on the problem.

Disaggregated Targets

Disaggregated targets may be of interest in:

(a) certain descriptive surveys,

(b) analytical surveys where the X_i may be viewed as 'background variables'.

In more detail:

(a) O'Muircheartaigh (1977) writes:

'The first and simplest purpose of multivariate analysis may be data description or data reduction. The aim in this case is to reduce the volume of data by transforming the full data set into a more compact form which preserves its essential characteristics and which provides an accurate summary.'

If differences between, say, within-cluster parameters exist, then we might view these as essential characterisitcs of the data which we wish to preserve by performing a disaggregated analysis. For example, Holt et al (1976) consider the correlations between educational tests and attitude variables from a survey of schoolchildren, where schools are clusters. They find that the correlation structure differs between schools and they suggest that it is more illuminating to investigate correlations within certain school types, of similar correlation structure, than to consider a single aggregate correlation matrix.

(b) The idea of disaggregating a population into subgroups, with their implied social structure, for the purpose of causal analysis, features widely in the social sciences (for example Dogan and Rokkam 1969 or, with special reference to the survey context, Coleman, 1959). Galtung (1967, pp.37-38) writes:

'A unit may be seen, judged and measured not only in absolute terms but also relative to other units of the same kind belonging to the same set. And it may often be fruitful to look for the *structure* of the set. Secondly, it often happens that the set of units itself is a unit of analytical interest and this unit itself may again generate a set of interest in some context'.

In the discussion of Kish and Frankel (1974), Kalton, Sampford and Brown question the value of regressing across strata when different regression relationships hold in the different strata. The value of regressing across clusters may similarly be questioned. The case for a disaggregated analysis is particularly strong when there are prior substantive reasons for suspecting that subgroup membership influences the regression relationship, as say when the subgroups are institutions or countries.

For example, consider a survey of perinatal mortality among babies born in hospitals where the clusters are hospitals. Suppose the dependent variable Y_1 is a mortality rate and the independent variable Y_2 is the distance from the mother's home to the hospital. We might be interested in whether there is a greater mortality amongst babies born of mothers living further from hospitals and the consequent policy implications for siting of hospitals. Now Y_2 will on average be greater in rural hospitals than in urban hospitals. In Figure 1.1 we represent the situation where a distance effect exists in both types of hospital but where the overall regression slope is zero. In Figure 1.2 there is no distance effect in either type of hospital but overall a positive effect exists. In each case we suggest the relevant policy implications are derived from the disaggregated within cluster regressions (the different levels in mortality for urban and rural hospitals would, of course, also have other policy implications).

Hypothetical regression relationships between perinatal mortality, Y_1 and distance from hospital, Y_2 for rural and urban hospitals.

Figure 1.1

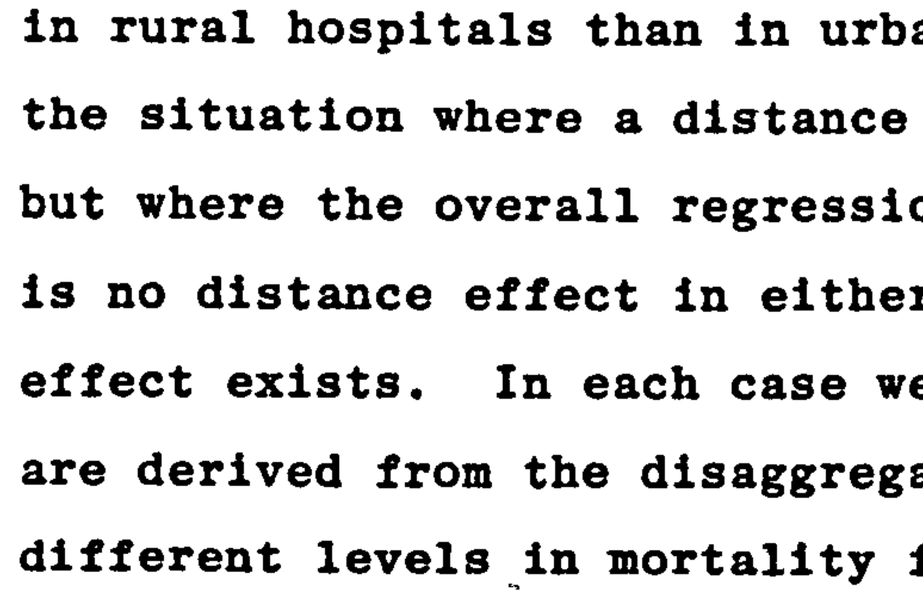
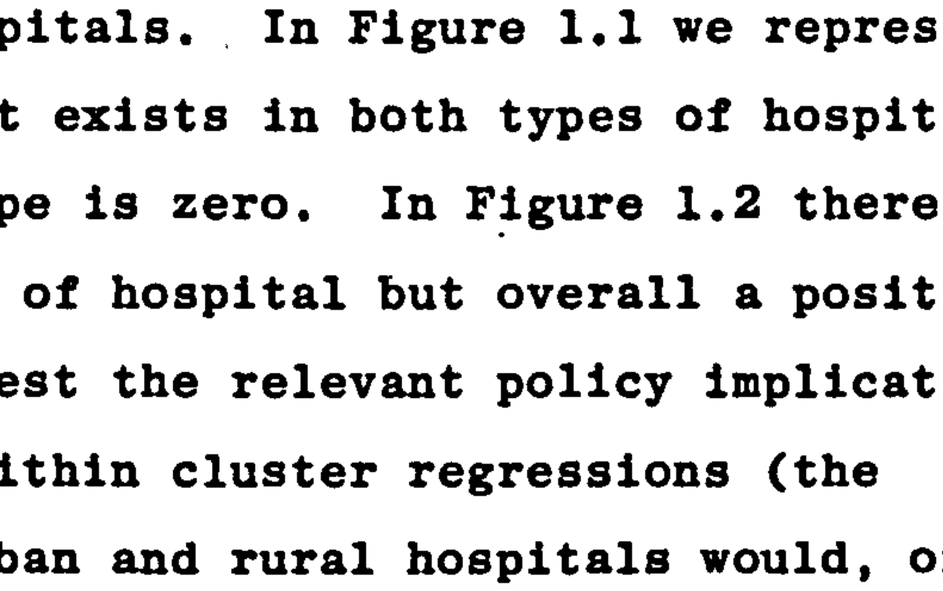


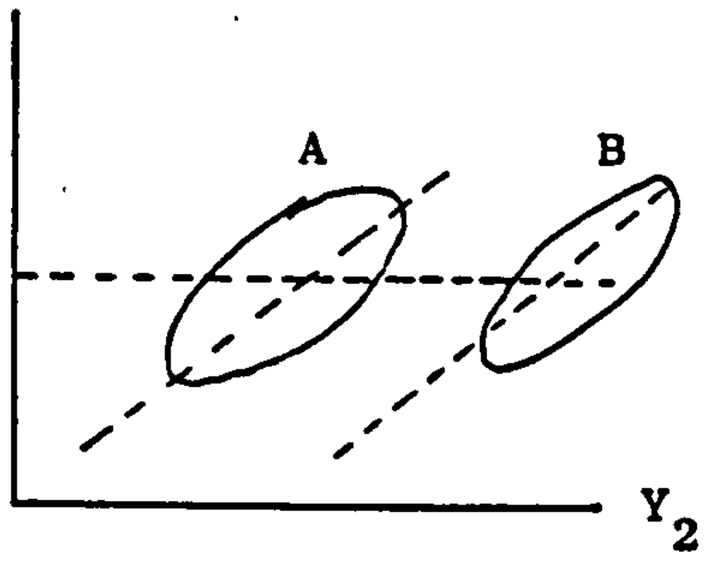
Figure 1.2



Similar arguments would apply to surveys of schoolchildren, where schools are used as subgroups. Relationships between variables typically may depend on the school environment (e.g. Rutter et al, 1979) and so it may be appropriate to examine disaggregated relationships within schools and then separately to compare relationships between schools.

In the above two institutional examples there are prior reasons for expecting the subgroup relationships to be directly influenced by subgroup membership. However, in many surveys strata or clusters form arbitrarily defined geographically continguous areas which possess no well-defined causal interpretation. Yet even in this case it may be sensible to examine disaggregated relationships, since by so doing we may be controlling for 'extraneous variation' which may be desirable given the non-experimental design of a survey, (e.g. Fields, 1971; Bielby, 1981). For example suppose we wish to regress Y_1, household expenditure on a given commodity, on Y_2 household income. Suppose one stratum is in an area containing predominately pensioner households with low incomes but small household size. Suppose another stratum is an area containing mainly young families with higher average household income and with larger household sizes. Then a hypothetical regression relationship in each stratum is represented in Figure 1.3. Again for analytical purposes it may be sensible to examine within subgroup relationships if the effect of increasing income is of interest.

Figure 1.3: Hypothetical Relationship between Y_2, income and Y_1 expenditure



To summarise and generalise, suppose that we may split $\underline{Y}$ into $(\underline{Y}_1\ \underline{Y}_2)$ where $\underline{Y}_1$ are the endogenous variables and $\underline{Y}_2$ are the exogenous variables. We have argued that if the X variables may be viewed as 'background variables' then we may write $p(\underline{Y}|\underline{X}, \theta) = p(\underline{Y}_1|\underline{Y}_2, \underline{X}, \theta_1)$ $P(\underline{Y}_2|\underline{X}, \theta_2)$ where $\theta = (\theta_1, \theta_2)$ and that θ_1 should be the object of inference. An argument against this approach is that the competent researcher should be able to specify $\underline{Y}_2$ correctly such that $\underline{Y}_1$ is conditionally independent of $\underline{X}$ given $\underline{Y}_2$ and hence, for example, the great effort involved in examining within cluster and within stratum regressions and intra-cluster correlations is unnecessary. According to this argument, all we have demonstrated above is an 'omitted variables' problem. If we had included dummy variables for urban/rural differences, for schools and for household type then 'design effects' would disappear. The basic argument against this case is that it is never possible to correctly specify $\underline{Y}_2$. For example, Mkai (1981) attempts to specify better and better models for a consumption function with an increasing number of independent variables using FES data and he shows that 'design effects' do not tend to disappear. Consider just the problem of residual intra-cluster correlations which one might argue is due to omitted variables. This is very similar to the problem of serially correlated residuals in the econometric analysis of time series. Economists do not like such serial correlation because it suggests omitted variables (but see Hendry and Mizon, 1978, for another explanation) but they are still prepared to fit models allowing for it because otherwise they will obtain inconsistent estimates of the remaining structural parameters.

Aggregate Targets

Aggregate targets of inference may be of interest in:

(a) certain descriptive surveys,

(b) analytical surveys where cluster membership may not be viewed as a background variable.

In more detail:

(a) Often aggregate characteristics of a population, such as the difference in mean income between men and women (which may be

viewed, in multivariate terms, as the linear regression coefficient of income on a dummy sex variable), will be of interest for descriptive purposes without attempting any causal interpretation. Sometimes, as in this example, a disaggregated analysis might also be of interest. However, on other occasions a disaggregated analysis might be irrelevant or even misleading. For example, suppose we wish to identify the 'components of mathematical ability' in students attending mathematics courses in U.K. universities. We take a two stage sample of students using universities as first-stage units. Suppose, hypothethically, that in pre-university examinations the students' pure mathematics ability Y_1 and applied mathematics ability Y_2 are measured perfectly and that the i^{th} university accepts a student if $Y_1 + Y_2 \geq k_i$. If, again hypothetically, students choose the university which would accept them with the highest k_i then the distribution of Y_1 and Y_2 within universities might resemble the 'slices' in Figure 1.4.

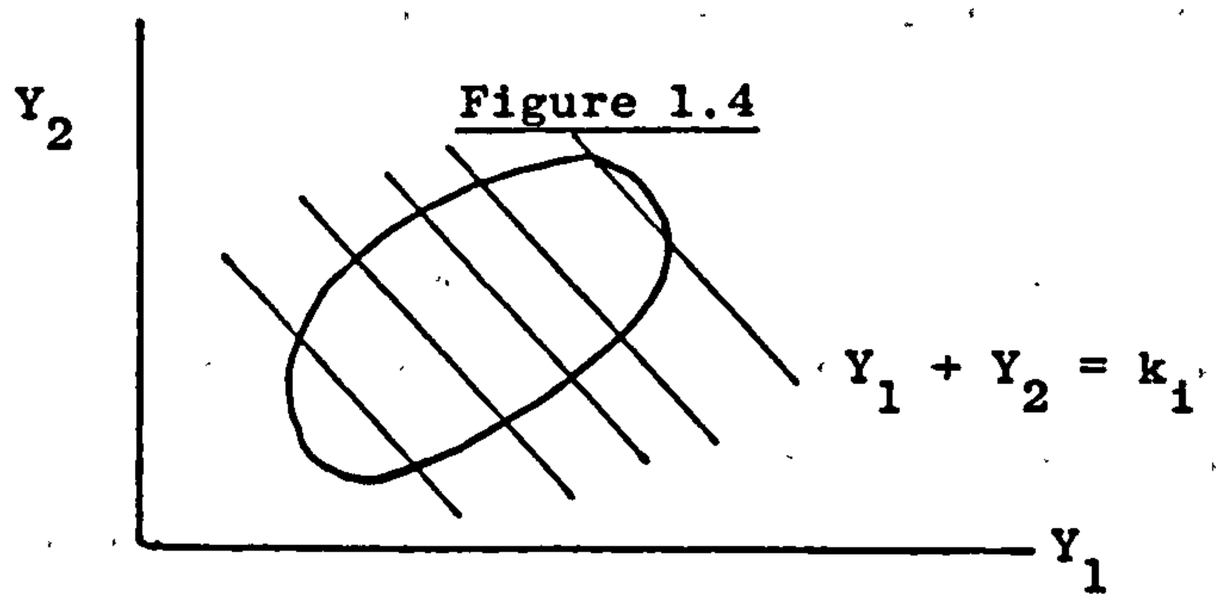


Figure 1.4

A within-university analysis might suggest zero or negative correlations between Y_1 and Y_2 and also suggest that the first principal component of ability was in the pure-applied direction rather than in the low ability-high ability direction. In this example the aggregate parameter might be a more meaningful object for inference.

(b) Consider a study of the determinants of income based on area cluster sample survey data. Cluster membership may itself be partly determined by income because of selective migration (e.g. Blalock, 1968) and a disaggregated regression analysis might underestimate the effects of independent variables. At the extreme, if income is constant within clusters, we would conclude that no variables

have any effect on income, *ceteris paribus*. Clearly it would be inappropriate to condition on cluster membership in this case. Goldberger (1981) distinguishes between *explicit* and *incidental selection* on the dependent variable. In sample surveys cluster membership would generally not be a perfect function of the dependent variable (explicit selection) and so we may refer to this case as one of incidental selection on the dependent variable.

Conclusion

Question (A) Section 1.1 might be rephrased: what are the properties of methods which ignore $p(s|\underline{x})$ and $\underline{x}$? If the true objects of inference are the disaggregated parameters then the answer to question (A) is that such methods are meaningless. For example, the fact that the aggregate regression coefficient might be zero in Figure 1.1 may have no causal relevance. Hence question (A) is really only interesting if the object of inference is an aggregate parameter.

Question (B) (what methods should be used) is meaningful for both disaggregated and aggregated targets. There are two advantages of disaggregated targets for this question. Firstly inference using the Sampling Theory Approach may be independent of the sample design (see Section 1.2.3). Secondly the problem is essentially a case of the conventional structured model of Section 1.1. In general standard methods may be applied, although some new considerations do arise in the case of clusters (e.g. Pfefferman and Nathan, 1981). The main disadvantage of dissaggregated targets is that researchers are unlikely to wish to define their topics of interest in terms of the vagaries of survey design. This is particularly so for descriptive purposes, but even for analytical purposes in regression analysis,say,it seems likely that researchers would rather select sufficient interpretable independent variables so that the regression function does not depend on $\underline{x}$ (even though the error structure may e.g. Campbell 1977, Sedransk 1966). In this case the disaggregated parameter is equivalent to the aggregate parameter ψ_o,

defined above,anyway.

In this thesis we shall be concerned only with aggregate targets and the main emphasis will be on description. This is just so as to define a manageable area for study and not to deny the importance of disaggregated targets.

1.4 Outline of Thesis

In this thesis we shall be concerned with two basic models/ selection schemes:

1. Pearson-type Model/selection section

This scheme is described in Section 2.1. A fundamental assumption of the model is that different units are independent.

2. Two-stage Model /selection scheme

This scheme is described in Section 5.1. The model allows for intra-cluster correlation.

For each of these models we shall define parameters of interest in terms of either

1. First and Second Order Moments

viz. means, variances and covariances.

or

2. Multivariate Methods

viz. correlation coefficients, regression analysis, principal components analysis, factor analysis.

In each case we shall only be interested in the point estimation of parameters and we shall ask two questions (see Sections 1.1 and 1.2.1).

(A) What are the properties of the standard (IID model/srs design) point estimators under the above model/selection schemes?

(B) What alternative estimators, whether design-based or model-based, might be adopted?

These three classifications allow us to define a 2^3 layout in which we may place Chapters 2-7 of this thesis.

	Pearson-type		Two-stage	
	Q.(A)	Q.(B)	Q.(A)	Q.(B)
1st and 2nd moments	Ch.2	Ch.3	Ch.5	Ch.6
Multi-variate methods	Ch.4		Ch.7	

There is also a fourth classification, finite or superpopulation parameters, which will divide the separate chapters 2-7, and which essentially defines a 2^4 layout. Chapter 8 provides a conclusion.

1.5 Review of Literature

The problem of the multivariate analysis of sample survey data as formulated in this Chapter has received almost no *direct* attention in the literature. The only major reference is Bebbington and Smith (1977) which investigates question (A) partly theoretically and mainly using a simulation study. The most related fields are (i) variance estimation of complex statistics (e.g. Kalton, 1977; Shah, 1978) and (ii) regression analysis of sample survey data (e.g. Smith, 1982). The first field has very limited relevance to this thesis because we are only concerned with point estimation. The second field is very related but is only a subsection of multivariate analysis. As such, it does not seem worthwhile to attempt an overall review of the literature discussing individual work. Instead we prefer to discuss individual papers in the separate chapters.

We have noted in Section 1.4. that this thesis falls into two basic parts : Part 1, Chapters 2-4, is based on a Pearson-type model/ selection scheme; Part II, Chapters 5-7, is based on a clustered model/selection scheme. Part I has evolved from some very specific work. Scott (1977) proposed a simple framework for viewing sample selection from a multivariate population. Smith (1978) made extra multivariate normality assumptions and derived maximum likelihood estimators. This approach is very attractive for our purposes because it may be naturally related to classical multivariate analysis. Nathan and Holt (1980), Holt et al (1980) and Smith (1982) used this framework to investigate regression analysis. In Part I we shall use the same framework to investigate multivariate analysis. Part II has less specific antecedants. The selection scheme is that of classical two-stage sampling. The model is a generalisation of the conventional random effects model, e.g. Scott and Smith (1969).

Although this thesis is related to work on sample selection in psychometrics, econometrics and biometrics, its main reference points are in the literature on the analysis of sample surveys. We therefore propose to provide an annotated bibliography of this subject, if only in order to give a rough quantitative measure of the importance attached

to various aspects of this subject by previous workers in the field. A review of some of this literature is given by Rao (1975).

The following classification is used (c.f. Section 1.4.):

Design:	MS - multi-stage
	ST - stratified
	Oth - other
Question : (see Section 1.4)	(A)
	(B)
Object of Inference:	Sup - superpopulation parameter
	Fin - finite population parameter
Inference :	Est - estimation
	HT - hypothesis testing

We need to make some caveats : (i) it is often very difficult to allocate papers to the above categories, for example papers on the analysis of categorical data from stratified samples may make no distinction between superpopulation and finite population inference, (ii) whereas we attempt to be 'largely' comprehensive some minor papers, unpublished papers and papers of only marginal relevance have been excluded. It is difficult to define 'analytical' surveys and there is of course a blurring of the edges between 'analytical' and 'descriptive'. We have omitted work on the estimation of totals, means, proportions, quantiles, domain means and ratios. We have also omitted work on variance estimation of complex statistics.

	Designs			Question		Object		Inference	
	MS	ST	Oth	(A)	(B)	Sup	Fin	Est	HT
Analysis of Categorical Data									
Adhikari and Sarma (1978)	*				*	*			*
Altham (1976)	*			*	*	*			*
Brier (1980)	*			*	*	*			*
Cohen (1976)	*			*	*	*			*

	Designs			Question		Object		Inference	
	MS	ST	Oth	(A)	(B)	Sup	Fin	Est	HT
Cowan & Binder (1978)	*			*			*		*
Fellegi (1980)	*	*		*	*		*		*
Freeman and Koch (1976)	*	*	*		*		*	*	
Holt et al (1980a)	*	*	*	*	*	*			*
Imrey et al (1979)		*			*		*	*	
Imrey et al (1980)	*	*	*	*	*		*	*	*
Lepkowski and Landis (1980)	*			*			*	*	
Nathan (1969)		*			*		*		*
Nathan (1972)		*			*		*		*
Nathan (1975)	*	*			*		*		*
Rao and Scott (1981)	*	*	*	*	*	*	*		*
Shuster and Downing (1976)	*	*			*		*	*	*
Tomberlin (1979)	*	*			*	*			*
Tomberlin (1980)	*	*			*	*		*	
Estimation of Differences(and linear combinations) of Domain Means									
Booth and Sedransk (1969)			*		*		*	*	
Frankel (1971)	*	*	*	*	*		*	*	
Freeman and Brock (1978)	*	*	*		*		*	*	*
Freeman et al (1976)	*	*	*		*	*		*	
Freeman et al (1977)	*	*	*		*		*	*	*
Kish (1969)	*	*	*	*	*		*	*	
Kish and Frankel (1970)	*	*	*	*	*		*	*	
Kish and Frankel (1974)	*	*	*	*	*		*	*	
Kish et al (1976)	*	*		*	*		*	*	
Koch et al (1975)	*	*	*		*	*	*	*	*
Koch and Lemeshow (1972)	*	*	*		*		*	*	
Liao and Sedransk (1975)			*		*		*	*	
Rao (1973)		*	*		*		*	*	
Sedransk (1965a)			*		*	*		*	
Sedransk (1965b)	*				*	*		*	*
Sedransk (1967)		*	*		*	*		*	*
Yates (1960)		*			*		*	*	
Regression									
Brewer & Mellor (1973)		*			*	*	*	*	
Campbell (1977)	*			*	*	*		*	

	Designs			Question		Object		Inference	
	MS	ST	Oth	(A)	(B)	Sup	Fin	Est	HT
Demets and Halperin (1977)			*	*	*	*		*	
Frankel (1971)	*	*	*	*	*		*	*	
Fuller (1975)	*	*			*		*	*	
Hartley and Silken (1975)		*	*	*	*		*	*	
Hill (1980)	*				*		*	*	
Holt and Scott (1981)	*	*		*	*	*		*	
Holt et al (1980b)		*	*	*	*	*		*	
Jonrup and Rennermalm (1976)			*		*		*	*	
Kish and Frankel (1970)	*	*	*	*	*		*	*	
Kish and Frankel (1974)	*	*	*	*	*		*	*	
Konijn(1962)	*				*	*		*	
Lemeshow (1977)		*			*		*	*	
Nathan and Holt (1980)		*	*	*	*		*	*	
Pfeffermann and Nathan (1981)	*				*	*		*	
Porter (1973)	*				*	*		*	
Shah et al (1977)	*	*		*	*		*		*
Smith (1982)	*	*	*	*	*	*	*	*	
Thomsen (1978)			*		*	*		*	
Estimation of (population) variance									
Chaudhuri (1978)			*		*		*	*	
Das and Tripathi (1978)			*		*		*	*	
Liu (1974a)			*		*		*	*	
Liu (1974b)			*		*		*	*	
Mukhopadhyay (1978)			*		*		*	*	
Wakimoto (1971a)	*				*	(*)	*	*	
Zacks (1981)			*		*		*	*	
Zacks and Solomon (1981)			*		*		*	*	
Estimation of covariances and correlation coefficients									
Aoyama (1954)	*				*	(*)	*	*	
Bebbington and Smith (1977)	*	*		*			*	*	
Frankel (1971)	*	*	*	*	*		*	*	
Gupta et al (1978)			*	*	*		*	*	
Gupta et al (1979)		*		*	*		*	*	
Kish and Frankel (1970)	*	*	*	*	*		*	*	
Kish and Frankel (1974)	*	*	*	*	*		*	*	

	Designs			Question		Object		Inference	
	MS	ST	Oth	(A)	(B)	Sup	Fin	Est	HT
Koop (1970)		*		*	*		*	*	
Wakimoto (1971b)		*			*	(*)	*	*	
Wakimoto (1971c)		*			*	(*)	*	*	
Principal Components Analysis									
Bebbington and Smith (1977)	*	*		*			*	*	
Tortora (1980)		*		*	*		(*)	(*)	

CHAPTER TWO - STANDARD ESTIMATORS UNDER PEARSON-TYPE SELECTION SCHEME

2.1 Framework

Recall the notation of Section 1.2. $U = \{1...N\}$ is a population of N identifiable units. Associated with the i^{th} unit of U is a pair of vectors (x_{1i}, x_{2i}) of dimensions (p_1, p_2) $(i=1...N)$. Let

$$\underline{x}_1 = (x'_{11} \ldots x'_{1N})' \tag{2.1}$$

$$\underline{x}_2 = (x'_{21} \ldots x'_{2N})' \tag{2.2}$$

We suppose that $\underline{x}_2$ is known and that a sample (subset), s, of U is selected according to a sampling design, $p(s|\underline{x}_2)$, which depends only on $\underline{x}_2$. x_{1i} is observed for $i \in s$ and unobserved for $i \notin s$.

According to the unconditional superpopulation model approach (Section 1.2.2) we assume that $(\underline{x}_1, \underline{x}_2)$ are joint realisations of the pair of random vectors, $(\underline{X}_1, \underline{X}_2)$, where

$$\underline{X}_1 = (X'_{11} \ldots X'_{1N})'$$

$$\underline{X}_2 = (X'_{21} \ldots X'_{2N})' \tag{2.3}$$

We assume that $\underline{X}_1$ and $\underline{s}$ are conditionally independent given $\underline{X}_2 = \underline{x}_2$.

N is fixed and known. We consider two specifications of the joint probability distribution of $(\underline{X}_1, \underline{X}_2)$ (c.f. Section 1.2.2).

Model I (the 'true' model)

The pairs (X_{1i}, X_{2i}) $(i=1...N)$ are mutually independent and identically distributed with joint probability density function $f(X_1|X_2)g(X_2)$ and with partitioned mean vector and covariance matrix

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \text{ respectively} \tag{2.4}$$

The distribution, $f(X_1|X_2)$, has the following properties (c.f. Pearson, 1912; Lawley, 1943):

$$\text{(i)} \quad E_I(X_1|X_2) = \mu_{1.2} + BX_2 \tag{2.5}$$

$$\text{where } \mu_{1.2} = \mu_1 - B\mu_2 \tag{2.6}$$

$$B = \Sigma_{12}\,\Sigma_{22}^{-1} \tag{2.7}$$

E_I is the expectation under Model I

i.e. the regression of X_1 on X_2 is linear,

$$\text{(ii)} \quad V_I(X_1|X_2) = \Sigma_{1.2} \tag{2.8}$$

$$\text{where } \Sigma_{1.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{2.9}$$

i.e. the covariance matrix of X_1 given X_2 is constant

Model II (the IID model)

The same assumptions are made as in Model I but also we assume that X_{1i} and X_{2i} are independent. Hence

$$E_{II}(X_1|X_2) = \mu_1 \tag{2.10}$$

$$V_{II}(X_1|X_2) = \Sigma_{11} \tag{2.11}$$

Notation

Without loss of generality let s = {1...n}.

Let $$f = n/N \tag{2.12}$$

$$\underline{x}_{1s} = (x_{11} \ldots x_{1n}) \;,\; \underline{X}_{1s} = (X_{11} \ldots X_{1n}) \tag{2.13}$$

$$\bar{x}_{1s} = \sum_{i\varepsilon s} x_{1i}/n \;,\; \bar{x}_{2s} = \sum_{i\varepsilon s} x_{2i}/n \tag{2.14}$$

$$\bar{x}_1 = \sum_{i\varepsilon U} x_{1i}/N \qquad \bar{x}_2 = \sum_{i\varepsilon U} x_{2i}/N \tag{2.15}$$

$$S_{11s} = \sum_{i\varepsilon s} (x_{1i} - \bar{x}_{1s})(x_{1i} - \bar{x}_{1s})'/(n-1) \tag{2.16}$$

$$S_{12s} = \sum_{i\varepsilon s} (x_{1i} - \bar{x}_{1s})(x_{2i} - \bar{x}_{2s})'/(n-1) \tag{2.17}$$

$$S_{22s} = \sum_{i\varepsilon s} (x_{2i} - \bar{x}_{2s})(x_{2i} - \bar{x}_{2s})'/(n-1) \tag{2.18}$$

$$S_{11} = \sum_{i\varepsilon U} (x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1)'/(N-1) \tag{2.19}$$

$$S_{12} = \sum_{i\varepsilon U} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)'/(N-1) \tag{2.20}$$

$$S_{22} = \sum_{i\varepsilon U} (x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)'/(N-1) \tag{2.21}$$

We shall make certain limiting arguments in this and the following two chapters. For this purpose we assume that $\underline{X}_1$ and $\underline{X}_2$ are the first N terms of infinite sequences of independent random variables $X_{11}, X_{12}...$ and $X_{21}, X_{22}...$ which follow either Model I or II. There will also be an infinite sequence of designs $p_N(s|\underline{x}_2)$. For simplicity we assume, as in Nathan and Holt (1980), that $p_N(s|\underline{x}_2)$ is of fixed size $n = n(N)$ and that $n \to \infty$ as $N \to \infty$. In order to distinguish between $O(n^{-1})$ and $O(N^{-1})$ say, we make no further assumption about the function n(N), e.g. Fuller (1975) assumes $n(N)/N = f$ converges to a limit in which case $O(n^{-1}) = O(N^{-1})$. We shall generally assume that sample moments, e.g. $\bar{x}_{2s}$ and S_{22s}, converge in probability as $n \to \infty$.

In this and the following chapter we take as twin objectives:

(i) the estimation of $\underline{\mu}_1$ and Σ_{11}, and

(ii) the prediction of x_1 and S_{11}.

In this chapter we consider the properties of the *standard IID estimators* $\bar{x}_{1s}$ and S_{11s} for both the above objectives. These estimators would be appropriate if Model II were correct. This chapter therefore involves an investigation of question (A) of Section 1.2.2., where we argued that differences between the properties of these estimators

under Models I and II may be interpreted as *effects* of *misspecifying* the model as Model II when in fact the true model is Model I. In Chapter 3 we shall address question (B) of Section 1.2.2 and consider alternative estimators of μ_1 and Σ_{11} and predictors of $\bar{x}_1$ and S_{11}.

In Section 1.2.2 we argued that question (A) was not strictly a problem of statistical inference, whereas question (B) was. We therefore leave a discussion of the problem of 'correct' inference until Section 3.1. In the remainder of this chapter we shall evaluate the properties of $\bar{x}_{1s}$ and S_{11s} as estimators of μ_1 and Σ_{11} or predictors of $\bar{x}_1$ and S_{11} in terms of

(i) 'sampling distributions' conditional on s and $\underline{x}_2$ i.e. in terms of the model distribution, f,

(ii) 'sampling distributions' conditional on $\underline{x}_2$ i.e. in terms of the model distribution, f, and the sampling design $p(s|\underline{x}_2)$,

(iii) 'sampling distributions' unconditionally i.e. in terms of the model distributions, f and g, and the sampling design $p(s|\underline{x}_2)$.

2.2 Properties of the standard estimators, $\bar{x}_{1s}$ and S_{11s}

2.2.1 Properties conditional on s and $\underline{x}_2$

In Theorem 2.1 we give the first two moments of $\bar{x}_{1s}$ as an estimator of μ_1 under both Models I and II. In Corollary 2.2 we give the corresponding results for the prediction of $\bar{x}_1$.

Theorem 2.1

$$E_I(\bar{x}_{1s}|s,\underline{x}_2) = \mu_1 + B(\bar{x}_{2s} - \mu_2)$$

$$E_{II}(\bar{x}_{1s}|S,\underline{x}_2) = \mu_1$$

$$V_I(\bar{x}_{1s}|S,\underline{x}_2) = \Sigma_{1.2}/n$$

$$V_{II}(\bar{x}_{1s}|s,\underline{x}_2) = \Sigma_{11}/n$$

(where $\bar{x}_{1s}$, $\bar{x}_{2s}$, B, μ_1, μ_2, $\Sigma_{1.2}$, Σ_{11} are defined in (2.14), (2.7), (2.4) and (2.9))

Corollary 2.2

$$E_I(\bar{x}_{1s}-\bar{x}_1|s,\underline{x}_2) = B(\bar{x}_{2s}-\bar{x}_2)$$

$$E_{II}(\bar{x}_{1s}-\bar{x}_1|s,\underline{x}_2) = 0$$

$$V_I(\bar{x}_{1s}-\bar{x}_1|s,\underline{x}_2) = (1-f)\ \Sigma_{1.2}/n$$

$$V_{II}(\bar{x}_{1s}-\bar{x}_1|s,\underline{x}_2) = (1-f)\ \Sigma_{11}/n$$

(where f is defined in (2.12))

Proof of Theorem 2.1

$$E_I(\bar{x}_{1s}|s,\underline{x}_2) = \sum_s E_I(X_{1i}|x_{2i})/n$$

$$= \mu_1 + B(\bar{x}_{2s}-\mu_2) \text{ from (2.5), (2.6) and (2.14)}$$

$$V_I(\bar{x}_{1s}|s,\underline{x}_2) = \sum_s V_I(X_{1i}|x_{2i})/n^2 \text{ since } X_{11}..X_{1n} \text{ independent}$$

$$= \Sigma_{1.2}/n \text{ from (2.8)}$$

The moments for Model II follow by substituting B = 0 into the above formulae since Model II is a special case of Model I.

Proof of Corollary 2.2

$$(\bar{x}_{1s}-\bar{x}_1) = \sum_U w_i x_{1i} \quad \text{where } w_i = (1-f)/n \quad i=1...n \qquad (2.22)$$

$$= -1/N \qquad i=n+1...N$$

$$\sum_U w_i = 1-f-(N-n)/N = 0 \qquad (2.23)$$

$$\sum_U w_i^2 = (1-f)^2/n + (N-n)/N^2 = (1-f)/n \qquad (2.24)$$

Hence

$$E_I(\bar{x}_{1s}-\bar{x}_1|s,\underline{x}_2) = \sum_U w_i(\mu_{1.2}+Bx_{2i}) \quad \text{from (2.5) and (2.22)}$$

$$= B\ \Sigma w_i x_{2i} \quad \text{from (2.23)}$$

$$= B(\bar{x}_{2s}-\bar{x}_2)$$

$$V_I(\bar{x}_{1s}-\bar{x}_1|s,\underline{x}_2) = \Sigma\ w_i^2\ \Sigma_{1.2} \quad \text{from (2.8) and (2.22)}$$

$$= (1-f)\ \Sigma_{1.2}/n \quad \text{from (2.24)}$$

Again the results for Model II follow as a special case.

The main conclusion to be drawn from the above results is that model misspecification can introduce (conditional) *bias*. From Corollary 2.2 the (prediction) bias of $\bar{x}_{1s}$ is non-zero unless the sample is *balanced* on $\bar{x}_{2s}$ (e.g. Royall and Herson, 1973, a,b) and in general is 'linear' in $\bar{x}_{2s}$. There is a misspecification effect on the variance as well but this is $O_p(n^{-1})$ whereas the 'misspecification bias' is in general $O_p(1)$.

Similar conclusions apply to S_{11s} as an estimator of Σ_{11} or as a predictor of S_{11}. In order to show this we derive a general result which will also be used in Chapter 3.

Lemma 2.3

Let A be a random nxp matrix with rows which are independently distributed with common covariance matrix, Σ, but with possibly different means. Let M be a nxn symmetric matrix of constants and let $S = A'MA$. Then

$$E(S) = \bar{A}'M\bar{A} + tr(M)\Sigma$$

where $\bar{A} = E(A)$

Proof

Let the spectral decomposition of M be

$$M = \sum_{1}^{n} \lambda_i \gamma_i \gamma_i' \qquad (2.25)$$

Then
$$S = \Sigma \lambda_i y_i y_i' \qquad (2.26)$$

where
$$y_i = A'\gamma_i \qquad (2.27)$$

Now
$$E(y_i) = \bar{A}'\gamma_i \qquad (2.28)$$

and
$$\begin{aligned} \mathrm{cov}[(y_i)_k, (y_i)_\ell] &= \mathrm{cov}[\textstyle\sum_\alpha A_{\alpha k}(\gamma_i)_\alpha, \sum_\beta A_{\beta \ell}(\gamma_i)_\beta] \\ &= \textstyle\sum_\alpha (\gamma_i)_\alpha (\gamma_i)_\alpha \Sigma_{k\ell} \\ &= \Sigma_{k\ell} \end{aligned} \qquad (2.29)$$

where $(y_i)_k$ is the k^{th} element of y_i etc.

From (2.26), (2.28) and (2.29)

$$E(S) = \Sigma\ \lambda_i(\bar{A}'\gamma_i\gamma_i'\bar{A} + \Sigma)$$

$$= \bar{A}'\ M\ \bar{A} + tr(M)\Sigma \quad \text{from (2.25)}$$

The following theorem and corollary show that the bias of S_{11s} is in general non-zero. We shall only give the second moments of S_{11s} in the special case of normality (in Theorem 2.10) since under weak conditions (e.g. if the third and fourth moments of X_1 given X_2 are constant) the conditional covariance matrix of the elements of S_{11s} is $O_p(n^{-1})$ and hence, as for $\bar{x}_{1s}$, the major effect of misspecification is with respect to bias rather than second moments.

Theorem 2.4

$$E_I(S_{11s}|s,\underline{x}_2) = \Sigma_{11} + B(S_{22s} - \Sigma_{22})B'$$

$$E_{II}(S_{11s}|s,\underline{x}_2) = \Sigma_{11}$$

Corollary 2.5

$$E_I(S_{11s} - S_{11}|s, \underline{x}_2) = B(S_{22s} - S_{22})B'$$

$$E_{II}(S_{11s} - S_{11}|s,\underline{x}_2) = 0$$

Proof of Theorem 2.4

Let $A_1' = (x_{11} \ldots x_{1n})$ (2.30)

$A_2' = (x_{21} \ldots x_{2n})$ (2.31)

$P_w = I_n - 1_n 1_n'/n$ (2.32)

(where I_n is the nxn identity matrix, 1_n is the nx1 vector of ones)

Then $(n-1)S_{11s} = A_1'P_wA_1$ (2.33)

$(n-1)S_{22s} = A_2'P_wA_2$ (2.34)

The conditions of Lemma 2.3 apply under Model I, setting $S = (n-1)\ S_{11s}$, $A = A_1$, $M = P_w$, $\Sigma = \Sigma_{1.2}$. Since P_w is idempotent and of rank n-1 we have

$$tr(P_w) = n-1$$

And using (2.5) and the fact that $P_w 1_n = 0$ we have

$$P_w E_I(A_1 | s, \underline{x}_2) = P_w A_2 B' \qquad (2.35)$$

or in the notation of Lemma 2.3

$$P_w \bar{A}_1 = P_w A_2 B' \qquad (2.36)$$

Hence from Lemma 2.3

$$E_I((n-1)S_{11s} | s, \underline{x}_2) = \bar{A}_1' P_w \bar{A}_1 + (n-1)\ \Sigma_{1.2}$$

$$= (P_w \bar{A}_1)' P_w \bar{A}_1 + (n-1)\ \Sigma_{1.2} \text{ since } P_w \text{ is idempotent}$$

$$= BA_2'\ P_w A_2 B' + (n-1)\ \Sigma_{1.2} \text{ from (2.36)}$$

$$= (n-1)\left[BS_{22s}B' + \Sigma_{1.2}\right] \text{ from (2.34)}$$

Hence $E_I(S_{11s} | s, \underline{x}_2) = BS_{22s}B' + \Sigma_{1.2}$

$$= \Sigma_{11} + B(S_{22s} - \Sigma_{22})B' \text{ as required}$$

The result for Model II follows as a special case with B = 0.

Proof of Corollary 2.5

From Theorem 2.4 setting n = N

$$E_I(S_{11} | s, \underline{x}_2) = \Sigma_{11} + B(S_{22} - \Sigma_{22})B'$$

Hence from Theorem 2.4

$$E_I(S_{11s} - S_{11} | s, \underline{x}_2) = \Sigma_{11} + B(S_{22s} - \Sigma_{22})B' - \Sigma_{11} - B(S_{22} - \Sigma_{22})B'$$

$$= B(S_{22s} - S_{22})B' \text{ as required}$$

The result for Model II follows as a special case. Hence in general the effect of misspecification is to introduce (conditional) bias. The prediction bias of S_{11s} is non-zero unless the sample is balanced on S_{22s} and is in general 'linear' in S_{22s}.

A similar approach has been taken by Pearson (1903) which is the reason for the title of this chapter. He considered a multivariate normal population, P, for (X_1, X_2) and supposed that an infinite sub-population, P_A, was defined in terms of a set of values, A, of X_2.

$$P_A = \{\alpha \varepsilon P : X_2(\alpha) \varepsilon A\}$$

(a slightly more general version is given by Birnbaum et al, 1950).

He showed that the mean vector and covariance matrix of X_1 in P_A could be expressed in terms of the moments of X_2 in P_A as in Theorems 2.1 and 2.4. Aitkin (1934, 1935) expressed Pearson's (1903) results in matrix notation and Pearson (1912) and Lawley (1943) showed that the results also held under the less restrictive assumptions of Model I. Holt et al (1980b) discuss these results in the context of the regression analysis of sample surveys. If we let $n = N = \infty$ then $p(s|\underline{x}_2)$ defines a set of infinite subpopulations $P_{A(s)}$ of $P_{A(U)}$ a subpopulation of P. Pearson's results then apply in our framework conditional on s and $\underline{x}_2$.

Ledermann (1938a) also considered a multivariate normal population, P, for (X_1, X_2) and supposed that random samples, s, of size n were selected from P subject to the restriction that S_{22s} was fixed. His results are therefore formally identical to ours and he also derived Theorem 5.4 but by a rather long route.

We noted above that the 'misspecification biases' were of greater importance than the 'misspecification variances'. To emphasise this fact we might distinguish between *finite sample effects* and *selection effects*. The finite sample effects are reflected in the variability of the estimators as measured by their probability distribution given s and $\underline{x}_2$. Letting $n = N = \infty$ the finite sample effects disappear and we obtain Pearson's (1903) results:

$$\bar{x}_{1s} = \mu_1 + B(\bar{x}_{2s} - \mu_2) \qquad (2.37)$$

$$S_{11s} = \Sigma_{11} + B(S_{22s} - \Sigma_{22})\, B' \qquad (2.38)$$

The selection effects are then given by the asymptotic misspecification biases $B(\bar{x}_{2s} - \mu_2)$ and $B(S_{22s} - \Sigma_{22})B'$. Note that these selection effects essentially operate through the sampling design, $p(s|\underline{x}_2)$.

Geometric Approach

To aid the understanding of (2.38) we may adopt a geometrical construction. $(X_1)_1 \ldots (X_1)_{p1}\ (X_2)_1 \ldots (X_2)_{p2}$ define a basis for a

$(p_1 + p_2)$ - dimensional vector space $\mathcal{V}$ consisting of vectors of the form $\alpha'X_1 + \beta'X_2$ ($(X_1)_i$ is the i^{th} component of X_1 etc. and α and β are $p_1 \times 1$ and $p_2 \times 1$ vectors of scalars respectively). The covariance inner product (e.g. Dempster, 1969, p.269) may be defined on $\mathcal{V}$ as

$$< u, v > = cov_I(u, v)$$

Then vectors in $\mathcal{V}$ may be considered as Euclidean vectors in $R^{p_1+p_2}$ with lengths (norms) equal to their standard deviations (in the *population*) and where the cosine of the angle between two vectors is the correlation between the corresponding random variables. We assume that the $(X_1)_i$ and $(X_2)_j$ are linearly independent (i.e. the joint covariance matrix of (X_1, X_2) is of full rank). This geometric representation is more familiar at the sample level than at the population level where random variables are usually represented by orthogonal axes. At the sample level, n observations on p variables can be represented as p points in R^n with inner products about the mean being equal to the sample covariances. These p points are confined to a p-dimensional subspace of R^n and we might think of $\mathcal{V}$ as the 'limit' of this subspace as $n \to \infty$.

Let $R(X_2)$ be the range space of X_2, i.e. the subspace of $\mathcal{V}$ spanned by $(X_2)_1 \ldots (X_2)_{p2}$ and let $R^{\perp}(X_2)$ be the orthogonal complement of $R(X_2)$. Then the p_1 vectors $(X_1)_i$ may be written uniquely as

$$X_1 = (X_1 - BX_2) + BX_2$$

where the elements of $X_1 - BX_2$ lie in $R^{\perp}(X_2)$ and the elements of BX_2 lie in $R(X_2)$. $X_1 - BX_2$ is the orthogonal projection of X_1 on $R^{\perp}(X_2)$ and BX_2 is the orthogonal projection of X_1 onto $R(X_2)$.

We now consider a construction given for example by Thomson (1951, p.276) to account for selection.

Case 1: $p_2 = 1$

Define $X_2^* = \alpha X_2$ (2.39)

where $\alpha^2 = S_{22s}/\Sigma_{22}$

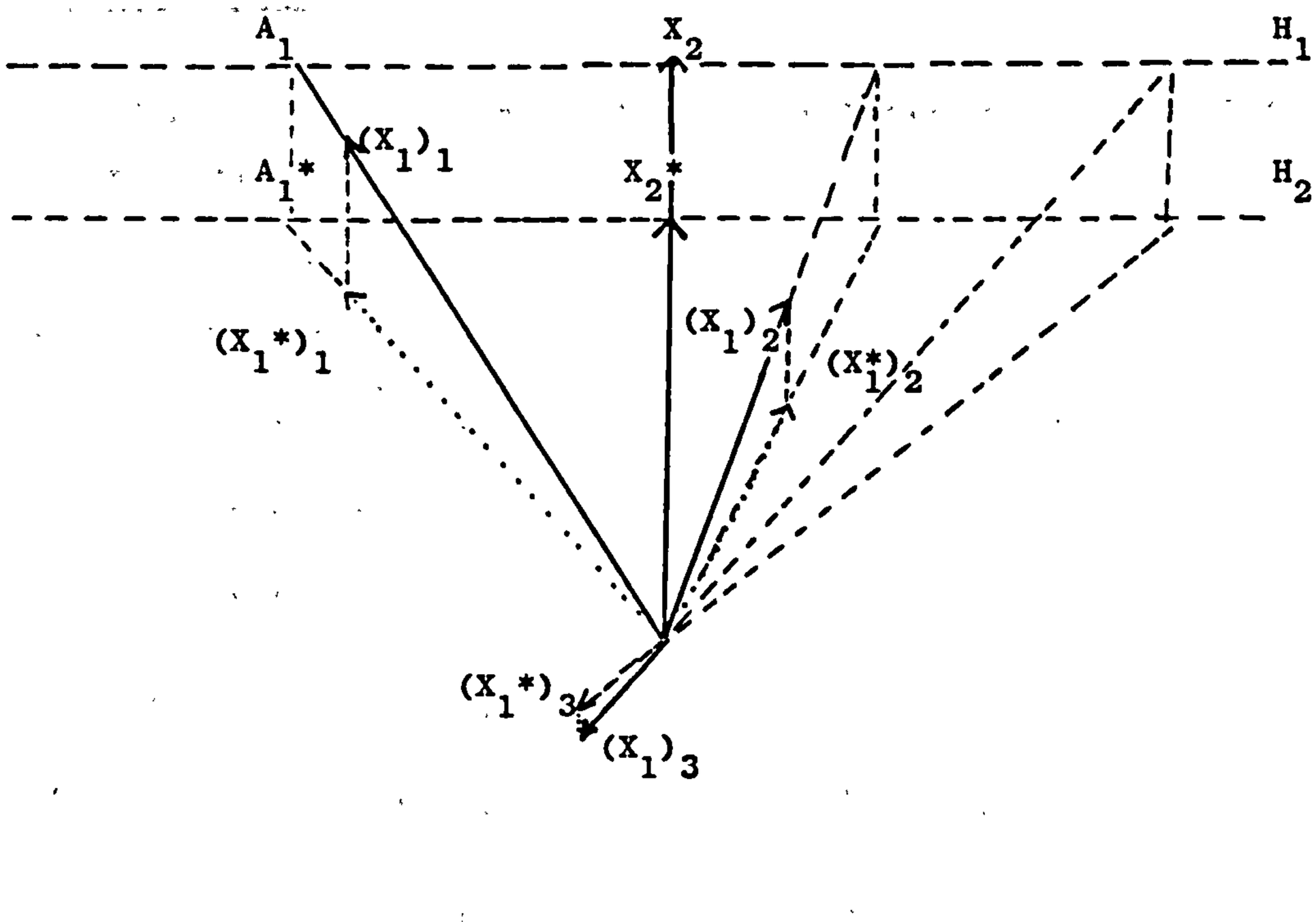
i.e. X_2^* is a vector in the same direction as X_2 but with squared length equal to the variance of X_2 in the *selected sample*.

Define $X_1^* = (X_1 - BX_2) + \alpha BX_2 = X_1 - (1-\alpha)\, BX_2$ (2.40)

i.e. X_1^* is the vector X_1 projected orthogonally towards $R^{\perp}(X_2)$ a proportion $(1-\alpha)$.

A geometrical construction of the vectors $(X_1^*)_i$ is indicated in Figure 2.1 by erecting the hyperplanes H_1 and H_2 through X_2 and X_2^* respectively orthogonal to $R(X_2)$. The inner products of the variables X_1^* and X_2^* in *the population* are now equal to the covariances between the corresponding random variables X_1 and X_2 in the *selected sample*, for

Figure 2.1



$$V_I(X_1^*) = \Sigma_{11} - (1-\alpha)\, B\, \Sigma_{21} - (1-\alpha)\, \Sigma_{12}B' + (1-\alpha)^2 B\Sigma_{22}B'$$

$$= \Sigma_{11} - 2(1-\alpha)\, \Sigma_{12}\Sigma_{22}^{-1}\, \Sigma_{21} + (1-\alpha)^2\, \Sigma_{12}\, \Sigma_{22}^{-1}\, \Sigma_{21}$$

$$= \Sigma_{11} - (1-\alpha^2)\, \Sigma_{12}\, \Sigma_{22}^{-1}\, \Sigma_{21}$$

$$= \Sigma_{11} + B(S_{22s} - \Sigma_{22})\, B'$$

$$= S_{11s} \quad \text{from (2.38)}$$

$$\text{cov}_I(X_1^*, X_2^*) = B\, \alpha^2\, \Sigma_{22} = B\, S_{22s}$$

$$V_I(X_2^*) = \alpha^2\, \Sigma_{22} = S_{22s}$$

Hence the effect of selection on the covariance matrix is the same as it would be if we reduced (or increased) each individuals' X_1 scores by a proportion of their X_2 score as in (2.40). Note that the selection effect on $(X_1)_i$ is zero if either $\alpha = 1$ or $(X_1)_i$ is orthogonal to X_2.

Case 2 : $p_2 > 1$

Our approach is similar to that of Ahmavaara (1954). Let A be a $p_2 \times p_2$ matrix such that

$$S_{22s} = A\, \Sigma_{22}\, A'$$

(e.g. if the spectral decompositions of S_{22s} and Σ_{22} are $S_{22s} = \Gamma_s \Lambda_s \Gamma_s'$, $\Sigma_{22} = \Gamma\, \Lambda\, \Gamma'$ then let $A = \Gamma_s \Lambda_s^{\frac{1}{2}}\, \Lambda^{-\frac{1}{2}} \Gamma'$)

$$\text{Let} \quad X_2^* = A\, X_2 \tag{2.41}$$

$$X_1^* = X_1 - BX_2 + BAX_2 = X_1 - B(I - A)\, X_2 \tag{2.42}$$

$$\text{Then } V_I(X_1^*) = \Sigma_{11} - \Sigma_{12}\, \Sigma_{22}^{-1}\, (I-A)\Sigma_{21} - \Sigma_{12}(I-A)'\, \Sigma_{22}^{-1}\, \Sigma_{21}$$

$$+ \Sigma_{12}\, \Sigma_{22}^{-1}\, (I-A)\Sigma_{22}(I-A)'\, \Sigma_{22}^{-1}\, \Sigma_{21}$$

$$= \Sigma_{11} + B\, (A\Sigma_{22}\, A' - \Sigma_{22})B'$$

$$= S_{11s}$$

Note that we may *also* make the mean of X_1^* equal to the mean of the selected sample in (2.27) if we let

$$X_1^* = X_1 - B(I-A)(X_2 - \mu_2) + B(\bar{x}_{2s} - \mu_2)$$

For the remainder of this section we obtain some distributional results under the assumption of normality.

Theorem 2.6

If X_1 has a multivariate normal distribution given X_2 then:

under Model I $\bar{x}_{1s}|s,\underline{x}_2 \sim N_{p_1}(\mu_1 + B(\bar{x}_{2s} - \mu_2), \Sigma_{1.2}/n)$

under Model II $\bar{x}_{1s}|s,\underline{x}_2 \sim N_{p_1}(\mu_1, \Sigma_{11}/n)$

Corollary 2.7

If X_1 has a multivariate normal distribution given X_2 then:

under Model I $\bar{x}_{1s} - \bar{x}_1|s, \underline{x}_2 \sim N_{p_1}(B(\bar{x}_{2s} - \bar{x}_2), (1-f)\Sigma_{1.2}/n)$

under Model II $\bar{x}_{1s} - \bar{x}_1|s,\underline{x}_2 \sim N_{p_1}(0, (1-f)\Sigma_{11}/n)$

Proof of Theorem 2.6 and Corollary 2.7

Both results follow from Theorem 2.1 and Corollary 2.2 because both $\bar{x}_{1s}$ and $\bar{x}_1$ are linear combinations of $x_{11}...x_{1N}$. Note that the distribution of $\bar{x}_{1s}$ depends on the design, $p(s|\underline{x}_2)$, *only* via $\bar{x}_{2s}$.

In Theorem 2.8 we give the distribution of S_{11s}. The distribution of $S_{11s} - S_{11}$ is complicated, being a weighted combination of non-central Wishart distributions, and is thus omitted.

Definition 2.1 : If $Z_1 \ldots Z_n$ are independent random p-vectors and $Z_i \sim N_p(\mu_i,\Sigma)$ (i=1...n) then $\Sigma Z_i Z_i'$ has a *non-central Wishart distribution*, denoted by $W_p(n, \Sigma, \tau)$, where $\tau = \Sigma \mu_i \mu_i'$ is the *non-centrality parameter* (c.f. Johnson and Kotz, 1972).

Theorem 2.8

If X_1 has a multivariate normal idstribution given X_2 then :

under Model I $(n-1)S_{11s}|s,\underline{x}_2 \sim W_{p_1}(n-1, \Sigma_{1.2}, (n-1)BS_{22s}B')$

under Model II $(n-1)S_{11s}|s,\underline{x}_2 \sim W_{p_1}(n-1, \Sigma_{11}, 0)$ (i.e. a central Wishart distribution)

Proof

Let the spectral decomposition of P_w (defined in (2.32)) be

$$P_w = \sum_{i=1}^{n-1} \gamma_i \gamma_i' \qquad (2.43)$$

(P_w is idempotent and of rank n-1 and so has (n-1) unit eigenvalues and one zero eigenvalue).

Then from (2.33) we may write, as in (2.26),

$$(n-1)\, S_{11s} = \sum_{1}^{n-1} y_i y_i'$$

where $y_i = A_1' \gamma_i$ (A_1 is defined in (2.30))

As in (2.29)

$$\begin{aligned} \mathrm{cov}_I\left[(y_i)_k, (y_j)_\ell \,|\, s, \underline{x}_2\right] &= \mathrm{cov}_I\left[\sum_\alpha A_{1\alpha k}(\gamma_i)_\alpha, \sum_\beta A_{1\beta\ell}(\gamma_j)_\beta \,|\, s, \underline{x}_2\right] \\ &= \sum_\alpha (\gamma_i)_\alpha (\gamma_j)_\beta \, \Sigma_{1.2k\ell} \\ &= \delta_{ij} \, \Sigma_{1.2k\ell} \end{aligned} \qquad (2.44)$$

where δ_{ij} is the Kronecker δ.

The y_i are linear combinations of $x_{11} \ldots x_{1n}$ and so are jointly normally distributed (given s and $\underline{x}_2$) and hence from (2.44) are independent (given s and $\underline{x}_2$). Let us write

$$y_i \,|\, s, \underline{x}_2 \sim N_{p_1}(\mu_i, \Sigma_{1.2})$$

Then from Definition 2.1.

$$(n-1)S_{11s} \sim W_{p_1}(n-1, \Sigma_{1.2}, \tau)$$

where $\tau = \Sigma \mu_i \mu_i'$

$$\begin{aligned} &= \sum_{i=1}^{n-1} E(A_1' \,|\, s, \underline{x}_2)\, \gamma_i \gamma_i'\, E(A_1 \,|\, s, \underline{x}_2) \\ &= E(A_1' \,|\, s, \underline{x}_2)\, P_w E(A_1 \,|\, s, \underline{x}_2) \quad \text{from (2.43)} \\ &= B A_2'\, P_w\, A_2\, B' \quad \text{from (2.35)} \\ &= (n-1)\, B S_{22s}\, B' \quad \text{from (2.34)} \end{aligned}$$

The distribution of S_{11s} under Model II follows as a special case. Note that the distribution of S_{11s} depends on the design only via S_{22s}. Note that even if $S_{22s} = \Sigma_{22}$ (i.e. balanced sampling) the distribution of S_{11s} is not the same under models I and II.

We now derive a general result which will also be used in Chapter 3.

Lemma 2.9

Under the conditions and notation of Lemma 2.3, if the rows of A have a multivariate normal distribution then

$$\text{cov}(S_{ij}, S_{k\ell}) = \text{tr}(M^2)(\Sigma_{ik}\Sigma_{j\ell}\Sigma_{jk}) + \Sigma_{j\ell}\,\psi_{ik} + \Sigma_{i\ell}\,\psi_{jk} + \Sigma_{jk} + \Sigma_{i\ell} + \Sigma_{ik}\;\psi_{j\ell}$$

where $\psi = \bar{A}'\, M^2\, \bar{A}$

Proof

From (2.26)

$$\text{cov}(S_{ij}, S_{k\ell}) = \text{cov}\left[\sum_\alpha \lambda_\alpha (y_\alpha)_i (y_\alpha)_j, \sum_\beta \lambda_\beta (y_\beta)_k (y_\beta)_\ell\right]$$

Since the rows of A have a multivariate normal distribution it follows by the argument in theproof of Theorem 2.8 that the y_α are independent. Hence

$$\text{cov}(S_{ij}, S_{k\ell}) = \sum_\alpha \lambda_\alpha^2 \,\text{cov}\left[(y_\alpha)_i\,(y_\alpha)_j,\ (y_\alpha)_k\,(y_\alpha)_\ell\right]$$

$$= \sum_\alpha \lambda_\alpha^2 \,\text{cov}\left[(\tilde{y}_\alpha)_i(\tilde{y}_\alpha)_j + (\tilde{y}_\alpha)_i(\mu_\alpha)_j + (\tilde{y}_\alpha)_j(\mu_\alpha)_i,\right.$$

$$\left.(\tilde{y}_\alpha)_k\,(\tilde{y}_\alpha)_\ell + (\tilde{y}_\alpha)_k\,(\mu_\alpha)_\ell + (\tilde{y}_\alpha)_\ell\,(\mu_\alpha)_k\right]$$

where $\mu_\alpha = E(y_\alpha)$

$$\tilde{y}_\alpha = y_\alpha - \mu_\alpha$$

The y_α (and hence $\tilde{y}_\alpha$) are normally distributed and so (e.g.

Anderson, 1958)

$$\text{cov}\left[(\tilde{y}_\alpha)_i(\tilde{y}_\alpha)_j,\ (\tilde{y}_\alpha)_k\ (\tilde{y}_\alpha)_\ell\right] = \Sigma_{ik}\ \Sigma_{j\ell} + \Sigma_{i\ell}\ \Sigma_{jk}$$

and $$\text{cov}\left[(\tilde{y}_\alpha)_i\ (\tilde{y}_\alpha)_j,\ (\tilde{y}_\alpha)_k\right] = 0$$

Hence

$$\text{cov}(S_{ij},S_{k\ell}) = \sum_\alpha \lambda_\alpha^2\ (\Sigma_{ik}\ \Sigma_{j\ell} + \Sigma_{i\ell}\ \Sigma_{jk} + \Sigma_{j\ell}\ \mu_{\alpha i}\ \mu_{\alpha k} + \Sigma_{jk}\ \mu_{\alpha i}\ \mu_{\alpha\ell} + \Sigma_{i\ell}\ \mu_{\alpha j}\ \mu_{\alpha k} + \Sigma_{ik}\ \mu_{\alpha j}\ \mu_{\alpha\ell})$$

but $\mu_\alpha = E(A'\gamma_\alpha) = \bar{A}'\gamma_\alpha$

$$\therefore\ \Sigma\lambda_\alpha^2\ \mu_{\alpha i}\ \mu_{\alpha j} = (\sum_\alpha \lambda_\alpha^2\ \bar{A}'\ \gamma_\alpha\ \gamma_\alpha'\ \bar{A})_{ij}$$

$$= (\bar{A}\ M^2\ \bar{A})_{ij}$$

$$= \psi_{ij}$$

and $\Sigma\ \lambda_\alpha{}^2 = \text{tr}\ (M^2)$

Hence

$$\text{cov}(S_{ij},S_{k\ell}) = \text{tr}(M^2)(\Sigma_{ik}\ \Sigma_{j\ell} + \Sigma_{i\ell}\ \Sigma_{jk}) + \Sigma_{j\ell}\ \psi_{ik} + \Sigma_{jk}\ \psi_{i\ell} + \Sigma_{i\ell}\ \psi_{jk} + \Sigma_{ik}\ \psi_{j\ell}$$ as required

We now use Lemma 2.9 to obtain the covariances between the elements of S_{11s} (which were omitted in Theorem 2 4) in the special case of normality.

Theorem 2.10

If X_1 has a multivariate normal distribution given X_2 then:

$$\text{cov}_I(S_{11sij},\ S_{11sk\ell}|s,\underline{x}_2) = (\Sigma_{1.2ik}\ \Sigma_{1.2j\ell} + \Sigma_{1.2i\ell}\ \Sigma_{1.2jk} + \Sigma_{1.2j\ell}\ \phi_{sik} + \Sigma_{1.2jk}\ \phi_{si\ell} + \Sigma_{1.2i\ell}\phi_{sjk} + \Sigma_{1.2ik}\ \phi_{sj\ell})/(n-1)$$

where $\phi_s = BS_{22s}B'$

$$\text{cov}_{II}(S_{11sij},\ S_{11sk\ell}|s,\underline{x}_2) = (\Sigma_{11ik}\ \Sigma_{11j\ell} + \Sigma_{11i\ell}\ \Sigma_{11jk})/(n-1)$$

Proof

Setting $S = (n-1)\ S_{11s}$, $A=A_1$, $M=P_w$, $\Sigma=\Sigma_{1.2}$ as in the proof of

Theorem 2.4 we obtain

$$tr(M^2) = tr(M) = (n-1)$$

$$\psi = \bar{A}_1' P_w^2 \bar{A}_1 = BA_2' P_w A_2 B' \quad \text{from (2.36)}$$

$$= (n-1)\, B\, S_{22s} B' \quad \text{from (2.34)}$$

$$= (n-1)\, \phi_s \qquad (2.45)$$

The result then follows by substituting into Theorem 2.9 (and noting that Model II is a special case of Model I with $\phi_s = 0$, $\Sigma_{1.2} = \Sigma_{11}$).

Corollary 2.11

Under the conditions of Theorem 2.10 we may write alternatively

$$cov_I(S_{11sij}, S_{11sk\ell} | s, \underline{x}_2) = (\Sigma^*_{ik}\, \Sigma^*_{j\ell} + \Sigma^*_{i\ell}\, \Sigma^*_{jk} - \phi_{sik}\, \phi_{sj\ell} - \phi_{si\ell}\, \phi_{sjk})/(n-1)$$

$$\text{where } \Sigma^* = \Sigma_{11} + B(S_{22s} - \Sigma_{22})B'$$

$$= E_I(S_{11s} | s, \underline{x}_2)$$

Proof

This follows by substituting $\Sigma_{1.2} = \Sigma^* - \phi_s$ into Theorem 2.10.

2.2.2 Properties conditional on $\underline{x}_2$

We now evaluate the properties of $\bar{x}_{1s}$ and S_{11s} over both the conditional model distribution of X_1 given X_2 and the randomisation distribution induced by the sampling design, $p(s|\underline{x}_2)$. Moments with respect to the latter distribution will be denoted by a subscript p. We assume that, under $p(s|\underline{x}_2)$, s is of fixed size n. In Section 2.2.1 we noted that the distributions of $\bar{x}_{1s}$ and S_{11s} given s and $\underline{x}_2$ under Model II did not depend on s. Hence the distributions are the same conditional on just $\underline{x}_2$ and we omit expressions for distributions under Model II in this section.

The following results are direct consequences of Theorem 2.1 and Corollary 2.2.

Theorem 2.12

$$E_{Ip}(\bar{x}_{1s}|\underline{x}_2) = \mu_1 + B(E_p(\bar{x}_{2s}|\underline{x}_2) - \mu_2)$$
$$V_{Ip}(\bar{x}_{1s}|\underline{x}_2) = \Sigma_{1.2}/n + B\ V_p(\bar{x}_{2s}|\underline{x}_2)B'$$

Corollary 2.13

$$E_{Ip}(\bar{x}_{1s} - \bar{x}_1|\underline{x}_2) = B(E_p(\bar{x}_{2s}|\underline{x}_2) - \bar{x}_2)$$
$$V_{Ip}(\bar{x}_{1s} - \bar{x}_1|\underline{x}_2) = (1-f)\ \Sigma_{1.2}/n + B\ V_p(\bar{x}_{2s}|\underline{x}_2)B'$$

For most practical designs $V_p(\bar{x}_{2s}|\underline{x}_2) = O_p(n^{-1})$ (Nathan and Holt, 1980). Hence, as in Section 2.2.1, the main effect of misspecification is the possible introduction of bias. If the design is epsem then from Corollary 2.13 the prediction bias of $\bar{x}_{1s}$ is zero i.e. the average of the model biases of $\bar{x}_{1s}$ over all samples s (given $\underline{x}_2$) is zero.

The following results are direct consequences of Theorem 2.3 and Corollary 2.4.

Theorem 2.14

$$E_{Ip}(S_{11s}|\underline{x}_2) = \Sigma_{11} + B(E_p(S_{22s}|\underline{x}_2) - \Sigma_{22})B'$$

Corollary 2.15

$$E_{Ip}(S_{11s} - S_{11}|\underline{x}_2) = B(E_p(S_{22s}|\underline{x}_2) - S_{22})B'$$

Hence if the first and second order selection probabilities are the same as srswor then the prediction bias of S_{11s} is zero. Again we may expect in general that the main effect of misspecification is in terms of bias.

We now consider two special cases.

Case 1: srswor

As noted above the prediction biases disappear in this case

$$E_{Ip}(\bar{x}_{1s} - \bar{x}_1 | \underline{x}_2) = 0$$

$$V_{Ip}(\bar{x}_{1s} - \bar{x}_1 | \underline{x}_2) = (1-f)(\Sigma_{11} + B(S_{22} - \Sigma_{22})B')/n$$

$$E_{Ip}(S_{11s} - S_{11} | \underline{x}_2) = 0$$

Note that even if we assume normality as in Theorem 2.10 the expression for $V_{Ip}(S_{11s} | \underline{x}_2)$ would be very complicated because of the term $V_p(S_{22s} | \underline{x}_2)$ (e.g. Hansen et al, 1953, p.101).

Case 2: stratified srswor

Using the notation of Example 1.1 we let

$$i \epsilon S_h \Leftrightarrow x_{2i} = e_h \quad i=1...N$$

Model I now implies that the X_{1i} are iid within strata with means $\mu_{1.2} + B_h$ and *common* covariance matrix $\Sigma_{1.2}$ where $B = (B_2...B_H)$, $B_1 = 0$.

Let $w_h = n_h/n$, $W_h = N_h/N$ (2.46)

$w' = (w_2...w_H)$, $W' = (W_2...W_H)$ (2.47)

Lemma 2.16

$$E_{Ip}(\bar{x}_{1s} - \bar{x}_1 | \underline{x}_2) = B(w-W)$$

$$V_{Ip}(\bar{x}_{1s} - \bar{x}_1 | \underline{x}_2) = (1-f)\ \Sigma_{1.2}/n$$

$$E_{Ip}(S_{11s} - S_{11} | \underline{x}_2) = B[n(\mathrm{diag}(w_h) - ww')/(n-1) - N(\mathrm{diag}\ (W_h) - WW')/(N-1)]B'$$

Proof

$$\bar{x}_{2s} = \sum_{i \epsilon S_h} n_h\, x_{2i}/n = w \qquad (2.48)$$

Similarly $\bar{x}_2 = W$ (2.49)

$$(n-1)(S_{22s})_{hk} = \sum_{\alpha} (x_{2\alpha})_h (x_{2\alpha})_k - n (\bar{x}_{2s})_h (\bar{x}_{2s})_k$$

$$= \delta_{hk} \sum_{\alpha \varepsilon S_h} 1 - n w_h w_k$$

$$= n_h \delta_{hk} - n w_h w_k$$

$$= n(w_h \delta_{hk} - w_h w_k)$$

$$\therefore (n-1) S_{22s} = n(\mathrm{diag}(w_h) - ww') \quad (2.50)$$

Similarly $(N-1)S_{22} = N(\mathrm{diag}(W_h) - WW')$ (2.51)

The results then follow by noting that w is fixed under $p(s|\underline{x}_2)$ and substituting into Corollaries 2.13 and 2.15.

Hence $\bar{x}_{1s}$ is an unbiased predictor of $\bar{x}_1$ under proportional allocation but not in general. Under proportional allocation the prediction bias of S_{11s} is $O(n^{-1})$. This result might be compared with that of Bebbington and Smith (1977) who consider the p-expectation of S_{11s} under this design. We may write

$$E_p(S_{11s}|\underline{x}_1, \underline{x}_2) = [\Sigma(n_h-1) S_{11h} + \Sigma n_h(\bar{x}_{1h} - \bar{x}_1^*)(\bar{x}_{1h} - \bar{x}_1^*)'$$

$$+ \Sigma(1-f_h)(1-W_h) S_{11h}]/(n-1)$$

where $S_{11h} = \sum_{S_h} (x_{1i} - \bar{x}_{1n})(x_{1i} - \bar{x}_{1h})'/(N_h-1)$

$$\bar{x}_{1h} = \sum_{S_h} x_{1i}/N_h$$

$$f_h = n_h/N_h$$

Bebbington and Smith (1977) make several approximations, one of which is to set the first coefficient of S_{11h} equal to n_h. This leads to the rather misleading conclusion that the bias of S_{11s} depends fundamentally

on H, the number of strata. They consider the case of proportional allocation when w = W. In this case the exact result is that

$$E_p(S_{11s}|\underline{x}_1, \underline{x}_2) = S_{11} + (N-n)(S_{11} - \Sigma W_h S_{11h})/N(n-1)$$

We may thenobtain

$$E_{Ip}(S_{11s} - S_{11}|\underline{x}_2) = (N-n)\ [\underline{B}(\text{diag}(W_h) - WW')B']/N(n-1)$$

i.e. the same as Lemma 2.16 in the case of proportional allocation.

2.2.3 Unconditional Properties

We now consider the properties of $\bar{x}_{1s}$ and S_{11s} evaluated over both the joint model distribution of (X_1, X_2) and the sampling design, $p(s|\underline{x}_2)$. Again we assume the sample size n is fixed and note that results for Model II are obtainable from Section 2.2.1.

The following results are direct consequences of Theorem 2.12 and Corollary 2.13.

Theorem 2.17

$$E_{Ip}(\bar{x}_{1s}) = \mu_1 + B(E_{Ip}(\bar{x}_{2s}) - \mu_2)$$

$$V_{Ip}(\bar{x}_{1s}) = \Sigma_{1.2}/n + B\ V_{Ip}(\bar{x}_{2s})\ B'$$

Corollary 2.18

$$E_{Ip}(\bar{x}_{1s} - \bar{x}_1) = B[E_{Ip}(\bar{x}_{2s}) - \mu_2]$$

$$V_{Ip}(\bar{x}_{1s} - \bar{x}_1) = (1-f)\ \Sigma_{1.2}/n + B\ V_{Ip}(\bar{x}_{2s} - \bar{x}_2)B'$$

Again for most practical sampling designs we may expect the variances to be of $O(n^{-1})$ and the misspecification bias to be of central importance. Note that $\bar{x}_{1s}$ is unconditionally unbiased for μ_1 or $\bar{x}_2$ under Model I if the first order inclusion probabilities $\pi_i(\underline{x}_2)$ do not depend on $\underline{x}_2$ e.g. in epsem designs of fixed size.

The following results are direct consequences of Theorem 2.14 and Corollary 2.15.

Theorem 2.19

$$E_{Ip}(S_{11s}) = \Sigma_{11} + B(E_{Ip}(S_{22s}) - \Sigma_{22})\ B'$$

Corollary 2.20

$$E_{Ip}(S_{11s} - S_{11}) = B(E_{Ip}(S_{22s}) - \Sigma_{22})\ B'$$

Again S_{11s} is unconditionally unbiased for Σ_{22} or S_{22} if the first and second order inclusion probabilities, $\pi_i(\underline{x}_2)$ and $\pi_{ij}(\underline{x}_2)$ do not depend on $\underline{x}_2$.

We now consider the two special designs of Section 2.2.2.

Case 1 : srswor

If the sample size n does not depend on $\underline{x}_2$ then $x_{21} \ldots x_{2n}$ are IID with mean μ_2 and covariance matrix Σ_{22}. Hence from Theorem 2.1.

$$E_{Ip}(\bar{x}_{1s}) = \mu_1$$

$$V_{Ip}(x_{1s}) = \Sigma_{11}/n$$

and from Corollary 2.2.

$$E_{Ip}(\bar{x}_{1s} - \bar{x}_1) = 0$$

$$V_{Ip}(\bar{x}_{1s} - \bar{x}_1) = (1-f)\Sigma_{11}/n$$

These are identical with the distributions under Model II as we would expect. Similarly, the same results will apply to S_{11s}.

Case 2 : stratified srswor

From Lemma 2.16 the moments of $\bar{x}_{1s}$ and S_{11s} depend on $\underline{x}_2$ via w and W. If $\underline{x}_2$ is distributed as in Example 1.3 then $N_1 \ldots N_H$ are multinomially distributed. For given designs, e.g. proportionate allocation, we could derive the joint distributio of (w,W) and hence obtain the unconditional moments of $\bar{x}_{1s}$ and S_{11s}. There seems little point in examining specific designs. Note that strata may also be defined by partitioning the sample space of X_2 into $A_1 \ldots A_H$ when

$$i \varepsilon S_h \Leftrightarrow x_{2i} \varepsilon A_h$$

The unconditional moments of $\bar{x}_{1s}$ and S_{11s} would be especially complicated in this case if the A_h depend on $\underline{x}_2$ e.g. Holt et al (1980) define the A_h in terms of quantiles of the realised values $x_{21} \ldots x_{2N}$.

2.3 Conclusions

If Model II were true then $\bar{x}_{1s}$ and S_{11s} would be natural estimators of μ_1 and μ_2 (or predictors of $\bar{x}_1$ and S_{11}) respectively. In this chapter we have investigated what the properties of these estimators would be if in fact Model I were true rather than Model II. The main effect of such misspecification is that of possible (asymptotic) bias. This may be in the sense of conditional bias given the actual sample s obtained and the finite population values $\underline{x}_2$ where the bias of $\bar{x}_{1s}$ is linear in $\bar{x}_{2s}$ and is zero only when $\bar{x}_{2s} = \bar{x}_2$ (balanced on the mean) or $\Sigma_{12} = 0$ and the bias of S_{11s} is linear in S_{22s} and is zero only when $S_{22s} = S_{22}$ (balanced on the covariance matrix) or $\Sigma_{12} = 0$. The averaged conditional bias of $\bar{x}_{1s}$ over all samples generated by $p(s|\underline{x}_2)$ will be zero if the design is epsem in which case the averaged conditional bias of S_{11s} is approximately zero (in general of $O(n^{-1})$. This suggests that for non-epsem designs there is a great need for considering alternative estimators, as in the next chapter, and even for unbalanced epsem designs there appears to be scope for alternative estimators which reduce or remove the conditional bias.

CHAPTER THREE - ALTERNATIVE ESTIMATORS UNDER PEARSON-TYPE SELECTION SCHEME

3.1 Introduction

In this chapter we consider alternative estimators of μ_1 and Σ_{11} and predictors of $\bar{x}_1$ and S_{11} to those discussed in Chapter 2. In Section 3.2 we consider estimators of μ_1 and Σ_{11} based on Model I of Section 2.1. In Section 3.3 we consider predictors of $\bar{x}_1$ and S_{11} based on Model I. In Section 3.4 we consider design-based predictors of $\bar{x}_1$ and S_{11}.

In Chapter 2 we addressed question (A) of Section 1.2.2 which we noted is not strictly a question of statistical inference. In this chapter we address question (B) of Section 1.2.2 which is a question of statistical inference. We shall adopt the Sampling Theory Approach to inference described in Section 1.2.3. This approach raises the 'problem of conditioning'. Are there ancillary statistics upon which we may condition when making inference about parameters of interest? In Chapter 2 we considered the sampling distribution of estimators condition on $\underline{x}_2$ or on $(s, \underline{x}_2)$. We might specifically ask therefore whether either of these statistics is ancillary for μ_1 and Σ_{11} or is predictive ancillary for $\bar{x}_1$ and S_{11}. In order to answer these questions we need to make distributional assumptions. In Example 1.3 (which was not strictly an instance of Model I since the variance of X_1 depended on X_2) both $\underline{x}_2$ and $(s, \underline{x}_2)$ would be predictive ancillary for $\bar{x}_1$ and S_{11} but neither would be ancillary for μ_1 or Σ_{11}. It is clear also from the arguments of Section 1.2.3 that if (X_1, X_2) were jointly multivariate normal then again both $\underline{x}_2$ and $(s, \underline{x}_2)$ would be predictive ancillary for $\bar{x}_1$ and S_{11} (provided $(\mu_{1.2}, B, \Sigma_{1.2})$ were Cartesian independent of (μ_2, Σ_{22})), but neither would be ancillary for μ_1 or Σ_{11}. The fact that we cannot strictly appeal to the Conditionality Principle for making inference about μ_1 or Σ_{11} is annoying, for example because we cannot relate the quality of an estimator to the quality of the selected sample, s. Since we can make μ_1 and Σ_{11} arbitrarily close to $\bar{x}_1$ and S_{11} respectively by increasing N it does seem unnecessarily formal not to condition on $(s, \underline{x}_2)$ and therefore, as noted in Section 1.2.3, we shall generally adopt a conditional approach in this chapter as in Holt et al (1980 b). This does, of course, make the mathematics neater.

3.2 Model-Based Estimation

In this section we assume that Model I of Section 2.1 is true. Maximum likelihood estimators (MLE's) of μ_1 and Σ_{11} are given in the following theorem (Smith, 1978; Smith, 1982) under the additional assumption of joint normality of (X_1, X_2).

Theorem 3.1

If Model I holds and the joint distribution of (X_1, X_2) is multivariate normal then the MLE's of μ_1 and Σ_{11} are:

$$\hat{\mu}_1 = \hat{\mu}_{1.2} + \hat{B}\,\bar{x}_2 = \bar{x}_{1s} + \hat{B}(\bar{x}_2 - \bar{x}_{2s}) \tag{3.1}$$

$$\hat{\Sigma}_{11} = \hat{\Sigma}_{1.2} + \hat{B}\,\tilde{S}_{22}\,\hat{B}' = S_{11s} + \hat{B}(\tilde{S}_{22} - \tilde{S}_{22s})\hat{B}' \tag{3.2}$$

where

$$\hat{\mu}_{1.2} = \bar{x}_{1s} - \hat{B}\,\bar{x}_{2s} \tag{3.3}$$

$$\hat{B} = S_{12s}S_{22s}^{-1} \tag{3.4}$$

$$\hat{\Sigma}_{1.2} = \tilde{S}_{11s} - \hat{B}\,\tilde{S}_{22s}\hat{B}' \tag{3.5}$$

$$\tilde{S}_{11s} = (n-1)S_{11s}/n\,,\ \tilde{S}_{22s} = (n-1)S_{22s}/n \tag{3.6}$$

$$\tilde{S}_{22} = (N-1)S_{22}/N \tag{3.7}$$

Proof

The likelihood may be expressed as in (1.2) as

$$p(\underline{x}_{1s}|\underline{x}_2, s, \mu_{1.2}, \Sigma_{1.2}, B)\ p(s|\underline{x}_2)\ p(\underline{x}_2|\mu_2, \Sigma_{22})$$

Eliminating $p(s|\underline{x}_2)$ which does not depend on the parameters, the likelihood reduces to that considered by Anderson (1957) who showed that $\hat{\mu}_1$ and $\hat{\Sigma}_{11}$ were as given in (3.1) and (3.2).

Note that $\hat{\mu}_1$ is the classical regression estimator of $\bar{x}_1$. It takes the naive estimator $\bar{x}_{1s}$ and adjusts it for the difference between the sample and population means of X_2 thus reducing the bias (see Theorems 2.1 and 3.3). Similarly, Σ_{11} may be viewed as a generalised regression estimator which adjusts S_{11s} for the difference between the sample and population covariances of X_2, thus reducing the bias (see Theorems 2.4 and 3.4). Note that if $n = N = \infty$

$$\hat{\mu}_1 = \bar{x}_{1s} + B(\mu_2 - \bar{x}_{2s})$$

$$\hat{\Sigma}_{11} = S_{11s} + B(\Sigma_{22} - S_{22s})B'$$

which are the solutions of equations (2.27) and (2.28) for μ_1 and Σ_{11} given by Pearson (1903).

Another representation of $\hat{\Sigma}_{11}$ (and $\hat{\mu}_1$) may be obtained by the geometrical approach discussed in Section 2.2.1. There we noted that, in Pearson's (1903) framework with $n = N = \infty$, the covariance matrix of X_1 in the *selected population* was the same as that of

$$X_1^* = X_1 - B(I-A)X_2 \tag{3.8}$$

(see 2.42) in the *superpopulation*. We now seek a random vector X_1^+ which has the same covariance matrix in the selected population as that of X_1 in the superpopulation, i.e. Σ_{11}. Now the covariance matrix of X_1^+ in the selected population is the same as the covariance matrix of

$$X_1^{+*} = X_1^+ - B^+(I-A)X_2 \tag{3.9}$$

(where B^+ is the regression coefficient matrix of X_1^+ on X_2) in the superpopulation. The natural choice for X_1^+ is obtained by setting $X_1^{+*} = X_1$ in (3.9) and solving for X_1^+

$$X_1^+ = X_1 + B^+(I-A)X_2$$

so that

$$B^+ = B + B^+(I-A)$$

or $B^+ = BA^{-1}$ assuming A is non-singular which it is if Σ_{22} and S_{22s} are positive definite

and so

$$X_1^+ = X_1 + BA^{-1}(I-A)X_2$$

We may represent X_1^+ geometrically as in Figure 2.1.

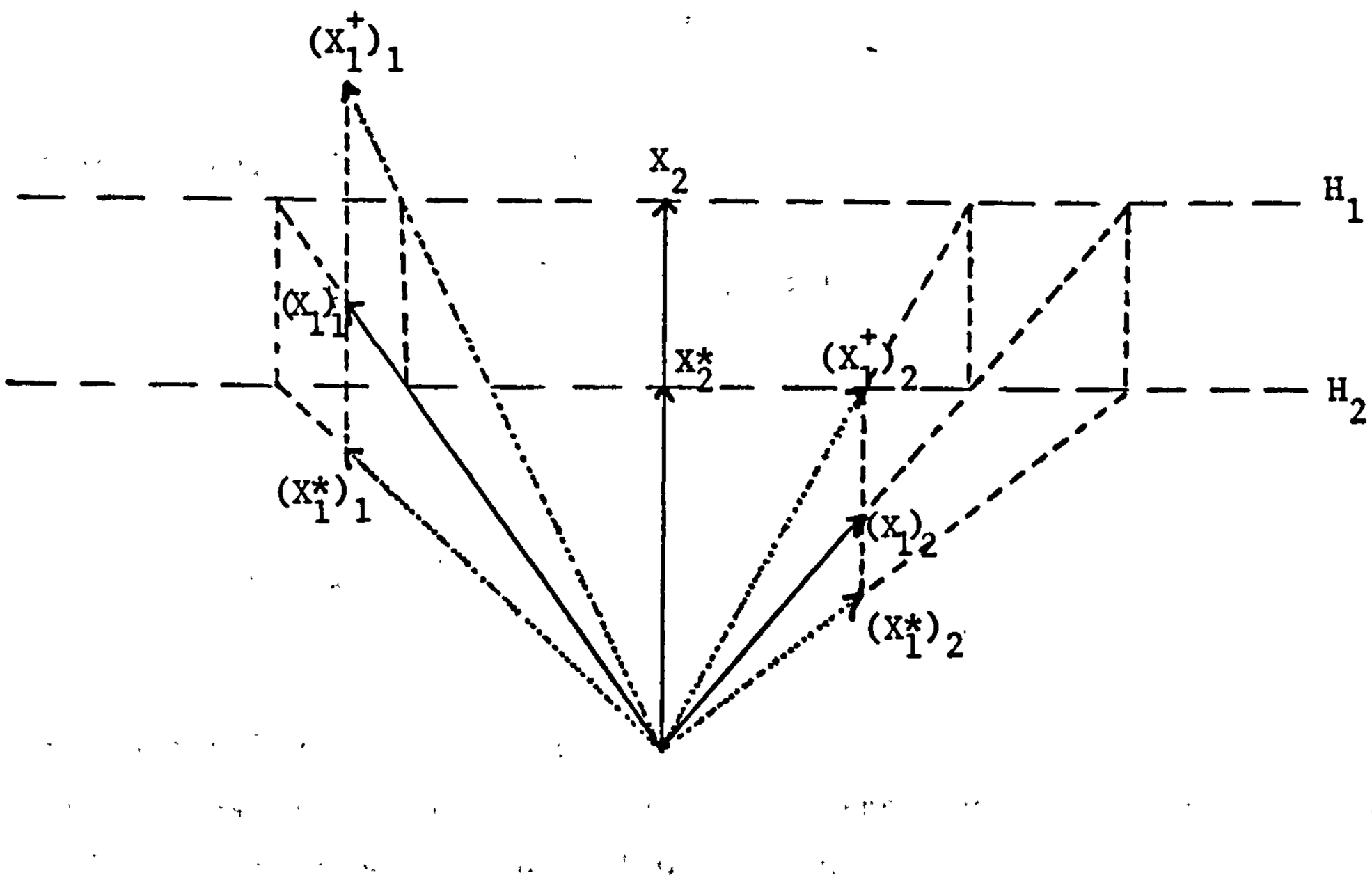


Figure 3.1

The obvious geometrical construction of X_1^+ is indicated in Figure 3.1. Whereas X_1 is projected onto X_1^* under selection, X_1^+ is projected onto $X_1 = X_1^{+*}$.

As in Section 2.2.1, we may by adding a constant term,define X_1^+ instead as

$$X_1^+ = X_1 + BA^{-1}(I-A)(X_2-\mu_2) - BA^{-1}\ (\bar{x}_{2s}-\mu_2) \qquad (3.10)$$

in which case in the selected population not only is the covariance matrix of X_1^+ equal to Σ_{11} but also the mean vector of X_1^+ is equal to μ_1 .

We now extend this approach to the case of finite n and N . Recalling that A is defined by

$$S_{22s} = A\Sigma_{22}A'$$

we propose to estimate A by $\hat{A}$ which is a $(p\times p)$ matrix such that

$$\tilde{S}_{22s} = \hat{A}\tilde{S}_{22}\hat{A}' \qquad (3.11)$$

Now define, following (3.10),

$$x_{1i}^+ = x_{1i} + \hat{B}\hat{A}^{-1}(I-\hat{A})(x_{2i}-\bar{x}_2) - \hat{B}\hat{A}^{-1}(\bar{x}_{2s}-\bar{x}_2) \qquad (3.12)$$

Let $\bar{x}_{1s}^+ = \sum_{i\varepsilon s} x_{1i}^+/n$

$$\tilde{S}_{11s}^+ = \sum_s (x_{1i}^+ - \bar{x}_{1s}^+)(x_{1i}^+ - \bar{x}_{1s}^+)'/n$$

Now according to the argument above, we would expect $\bar{x}_{1s}^+$ and $\tilde{S}_{11s}^+$ to be good estimators of μ_1 and Σ_{11} respectively *under selection*. This is confirmed by the following result:

Lemma 3.2

$$\bar{x}_{1s}^+ = \hat{\mu}_1$$

$$\tilde{S}_{11s}^+ = \hat{\Sigma}_{11}$$

Proof From (3.12)

$$\bar{x}_{1s}^+ = \bar{x}_{1s} + \hat{B}\hat{A}^{-1}(I-\hat{A})(\bar{x}_{2s}-\bar{x}_2) - \hat{B}\hat{A}^{-1}(\bar{x}_{2s}-\bar{x}_2)$$

$$= \bar{x}_{1s} - \hat{B}(\bar{x}_{2s}-\bar{x}_2)$$

$$= \hat{\mu}_1$$

$$\tilde{S}^{+}_{11s} = \tilde{S}_{11s} + \hat{B}\,\hat{A}^{-1}(I-\hat{A})\tilde{S}_{21s} + \tilde{S}_{12s}(I-\hat{A})'\hat{A}'^{-1}\hat{B}'$$

$$+ \hat{B}\,\hat{A}^{-1}(I-\hat{A})\tilde{S}_{22s}(I-\hat{A})'\hat{A}'^{-1}\hat{B}'$$

where $\tilde{S}_{12s} = (n-1)\,S_{12s}/n = \tilde{S}'_{21s}$

$$\therefore \quad \tilde{S}^{+}_{11s} = \tilde{S}_{11s} + \hat{B}[\hat{A}^{-1}S_{22s} - \tilde{S}_{22s} + \tilde{S}_{22s}\,\hat{A}'^{-1} - \tilde{S}_{22s} + \hat{A}^{-1}\tilde{S}_{22s}\hat{A}'^{-1}$$

$$- \hat{A}^{-1}\tilde{S}_{22s} - \tilde{S}_{22s}\,\hat{A}'^{-1} + \tilde{S}_{22s}]\hat{B}'$$

(since $\tilde{S}_{12s} = \hat{B}\,\tilde{S}_{22s}$, $\tilde{S}_{21s} = \tilde{S}_{22s}\,\hat{B}'$)

$$= \tilde{S}_{11s} + \hat{B}[\hat{A}^{-1}\tilde{S}_{22s}\,\hat{A}'^{-1} - \tilde{S}_{22s}]\hat{B}'$$

$$= \tilde{S}_{11s} + \hat{B}[\tilde{S}_{22} - \tilde{S}_{22s}]\hat{B}' \qquad \text{from (3.11)}$$

$$= \hat{\Sigma}_{11}$$

One application of this result might be in the use of standard 'IID-based' computer packages such as SPSS. The observations may be initially transformed by (3.12). Then the standard computed moments $\bar{x}^{+}_{1s}$ and $\tilde{S}^{+}_{11s}$ would be the MLE's and the observations would be asymptotically independent with common mean μ_1 and covariance matrix Σ_{11}. However, as we shall see the higher moments would not be the same as for an srs from the superpopulation. In a sense the x^{+}_{1i} are *model adjusted observations* as compared with conventional π-weighted observations. Unlike π-weighted observations the x^{+}_{1i} are in general functions not just of the x_{1i} but also the x_{1j} for $j \neq i$.

We now obtain the properties of $\hat{\mu}_1$ and $\hat{\Sigma}_{11}$ analogous to the results for $\bar{x}_{1s}$ and S_{11s} in Theorems 2.1 and 2.3. The distributions under Model II are of course no longer of interest.

Theorem 3.3

$$E_I(\hat{\mu}_1|s,x_2) = \mu_1 + B(\bar{x}_2 - \mu_2)$$

$$V_I(\hat{\mu}_1|s,\underline{x}_2) = \left(1 + (\bar{x}_{2s} - \bar{x}_2)'\,\tilde{S}^{-1}_{22s}(\bar{x}_{2s} - \bar{x}_2)\right)\Sigma_{1.2}/n$$

$$E_I(\hat{\Sigma}_{11}|s,\underline{x}_2) = \left[n - p_2 - 1 + tr(\tilde{S}_{22}\tilde{S}^{-1}_{22s})\right]\Sigma_{1.2}/n + B\tilde{S}_{22s}B'$$

Proof

$$\hat{\mu}_1 = \sum_s w_i x_{1i} \tag{3.13}$$

where
$$w_i = \left[1 - (x_{2i} - \bar{x}_{2s})' \tilde{S}_{22s}^{-1}(\bar{x}_{2s} - \bar{x}_2)\right]/n \tag{3.14}$$

Hence from (2.5)

$$E_I(\hat{\mu}_1|s,\underline{x}_2) = \Sigma\, w_i(\mu_{1.2} + Bx_{2i})$$

$$= \mu_{1.2} + B\bar{x}_{2s} - Bn\tilde{S}_{22s}\tilde{S}_{22s}^{-1}(\bar{x}_{2s} - \bar{x}_2)/n$$

$$= \mu_{1.2} + B\bar{x}_{2s} - B(\bar{x}_{2s} - \bar{x}_2)$$

$$= \mu_1 + B(\bar{x}_{2s} - \mu_2) \qquad \text{as required}$$

Similarly from (2.8) and (3.13)

$$V_I(\hat{\mu}_1|s,\underline{x}_2) = \sum_s w_i^2\, \Sigma_{1.2}$$

$$= \left[1 + \sum_s(\bar{x}_{2s} - \bar{x}_2)'\tilde{S}_{22s}^{-1}(x_{2i} - \bar{x}_{2s})(x_{2i} - \bar{x}_{2s})'\tilde{S}_{22s}^{-1}(\bar{x}_{2s}-\bar{x}_2)/n\right]\Sigma_{1.2}/n$$

$$= \left[1 + (\bar{x}_{2s} - \bar{x}_2)'\, \tilde{S}_{22s}^{-1}(\bar{x}_{2s} - \bar{x}_2)\right]\Sigma_{1.2}/n \qquad \text{as required}$$

From (3.2)

$$\hat{\Sigma}_{11} = \hat{\Sigma}_{1.2} + \hat{B}\, \tilde{S}_{22}\, \hat{B}' \tag{3.15}$$

We may write

$$n\, \hat{\Sigma}_{1.2} = A_1'\, M_1 A_1 \tag{3.16}$$

$$n\, \hat{B}\, \tilde{S}_{22}\, \hat{B}' = A_1'\, M_2\, A_1 \tag{3.17}$$

where
$$M_1 = P_W - P_W A_2(A_2'P_W A_2)^{-1}A_2'P_W \tag{3.18}$$

$$M_2 = nP_W A_2(A_2'P_W A_2)^{-1}\, \tilde{S}_{22}(A_2'P_W A_2)^{-1}A_2'P_W \tag{3.19}$$

and A_1, A_2 and P_W are defined in (2.30) - (2.32).

Now

$$tr(M_1) = tr(P_W) - tr(P_W A_2 (A_2' P_W A_2)^{-1} A_2' P_W)$$

$$= n-1 - tr\left(A_2' P_W A_2 (A_2' P_W A_2)^{-1}\right)$$

$$= n - p_2 - 1 \qquad (3.20)$$

and $$tr(M_2) = tr\left(n A_2' P_W A_2 (A_2' P_W A_2)^{-1} \tilde{S}_{22} (A_2' P_W A_2)^{-1}\right)$$

$$= tr(\tilde{S}_{22} \tilde{S}_{22s}^{-1}) \qquad (3.21)$$

since $$\tilde{S}_{22} = A_2' P_W A_2 / n$$

Hence from Lemma 2.3 and (2.36)

$$E(n \hat{\Sigma}_{1.2} | s, \underline{x}_2) = E(A_1' M_1 A_1 | s, \underline{x}_2)$$

$$= B A_2' M_1 A_2 B' + (n-p_2-1) \Sigma_{1.2} \qquad (3.22)$$

and $$E(n \hat{B} \tilde{S}_{22} \hat{B}' | s, \underline{x}_2) = E(A_1' M_2 A_1 | s, \underline{x}_2)$$

$$= B A_2' M_2 A_2 B' + tr(\tilde{S}_{22} \tilde{S}_{22s}^{-1}) \Sigma_{1.2} \qquad (3.23)$$

Now $$M_1 A_2 = 0 \qquad (3.24)$$

and $$A_2' M_2 A_2 = n \tilde{S}_{22} \qquad (3.25)$$

Hence from (3.15) and (3.23) - (3.25)

$$E(\hat{\Sigma}_{11} | s, \underline{x}_2) = \left[n - p_2 - 1 + tr(\tilde{S}_{22} \tilde{S}_{22s}^{-1})\right] \Sigma_{1.2}/n + B \tilde{S}_{22} B'$$

as required

Hence $\hat{\mu}_1$ has a bias of $O(N^{-1})$ compared with the bias of $\bar{x}_{1s}$ in Theorem 2.1 which was of $O(1)$. In order to consider the bias of $\hat{\Sigma}_{11}$ we may alternatively write:

$$E_I(\hat{\Sigma}_{11} | s, \underline{x}_2) = \Sigma_{11} + B(\tilde{S}_{22} - \Sigma_{22}) B' + \left[tr(\tilde{S}_{22} \tilde{S}_{22s}^{-1}) - p_2 - 1\right] \Sigma_{1.2}/n \qquad (3.26)$$

Provided $\mathrm{tr}(\tilde{S}_{22}\tilde{S}_{22s}^{-1})$ is of $O(1)$, which would seem reasonable for most practical sampling designs, the bias of $\hat{\Sigma}_{11}$ is $O(n^{-1})$, as we might expect given the usual bias of the MLE of the variance for an IID sample. The bias may be reduced by an ad hoc adjustment. For example, let

$$\hat{\hat{\Sigma}}_{11} = \lambda\hat{\Sigma}_{1.2} + \hat{B}\, S_{22}\, \hat{B}' \qquad (3.27)$$

where $\lambda = n[n - \mathrm{tr}(S_{22}S_{22s}^{-1}) - 1]/\,(n-p_2-1)(n-1)$

Then from (3.22) - (3.25)

$$E(\hat{\hat{\Sigma}}_{11}|\, s,\underline{x}_2) = \lambda(n-p_2-1)\Sigma_{1.2}/n + N[B\,\tilde{S}_{22}B' + \mathrm{tr}(\tilde{S}_{22}\tilde{S}_{22s}^{-1})\Sigma_{1.2}/n]/(N-1)$$

$$= [n - \mathrm{tr}(S_{22}S_{22s}^{-1}) - 1 + \mathrm{tr}(S_{22}S_{22s}^{-1})]\Sigma_{1.2}/(n-1) + BS_{22}B'$$

$$= \Sigma_{1.2} + BS_{22}B'$$

$$= \Sigma_{11} + B(S_{22} - \Sigma_{22})B' \qquad (3.28)$$

Hence the bias of $\hat{\hat{\Sigma}}_{11}$ is of $O(N^{-1})$.

The variance of $\hat{\mu}_1$ is of $O(n^{-1})$ as was that of $\bar{x}_{2s}$ in Theorem 2.1. However, the variance of $\hat{\mu}_1$ now dominates its bias in its MSE. The variance of $\hat{\Sigma}_{11}$ will also be of $O(n^{-1})$ under weak conditions and it will be given in the case of normality in Theorem 3.5.

We now consider the distributions of $\hat{\mu}_1$ and $\hat{\Sigma}_{11}$ under normality assumptions.

Theorem 3.4

If X_1 has a multivariate normal distribution given X_2 then under Model I

$$\hat{\mu}_1|\, s,\underline{x}_2 \sim N\left(\mu_1 + B(\bar{x}_2-\mu_2),\ [1 + (\bar{x}_{2s}-\bar{x}_2)'\tilde{S}_{22s}^{-1}(\bar{x}_{2s} - \bar{x}_2)]\Sigma_{1.2}/n\right)$$

Proof

From (3.13) $\hat{\mu}_1$ is a linear combination of independent normal random

variables and hence is normally distributed. The result follows from Theorem 3.3.

From (3.15) - (3.17) we may write

$$n\hat{\Sigma}_{11} = A_1'MA_1 \tag{3.29}$$

where
$$M = M_1 + M_2 \tag{3.30}$$

If the distribution of X_1 given X_2 is multivariate normal then the distribution of $\hat{\Sigma}_{11}$ given s and $\underline{x}_2$ will only be non-central Wishart if M is proportional to an idempotent matrix. But M_1 is idempotent and

$$M_1M_2 = M_2M_1 = 0 \tag{3.31}$$

Hence
$$M^2 = M_1 + M_2^2 \tag{3.32}$$

And so in general M is only idempotent if M_2 is idempotent. But M_2 is in general idempotent with probability zero, e.g. if $p_2 = 1$ then

$$M_2^2 = M_2\ \tilde{S}_{22}/\tilde{S}_{22s}$$

and M_2 is idempotent only if $\tilde{S}_{22s} = \tilde{S}_{22}$ or if $\tilde{S}_{22} = 0$. Hence in general the distribution of $\tilde{\Sigma}_{11}$ is more complicated than a non-central Wishart distribution, in fact it is a linear combination of independent non-central Wisharts. We can, however, give the covariances between the elements of $\hat{\Sigma}_{11}$ in the case of normality.

Theorem 3.5

If X_1 has a multivariate normal distribution given X_2 then

$$\begin{aligned}\mathrm{cov}_I(\hat{\Sigma}_{11ij}, \hat{\Sigma}_{11k\ell}|^s,\underline{x}_2) = {} & \left(n-p_2-1 + \mathrm{tr}(\tilde{S}_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}\tilde{S}_{22s}^{-1})\right) \\ & (\Sigma_{1.2ik}\Sigma_{1.2j\ell} + \Sigma_{1.2i\ell}\,\Sigma_{1.2jk})/n^2 \\ & + (\Sigma_{1.2j\ell}\psi_{ik}+\Sigma_{1.2jk}\psi_{i\ell}+\Sigma_{1.2i\ell}\psi_{jk}+\Sigma_{1.2ik}\psi_{j\ell})/n\end{aligned}$$

where $\psi = B\ \tilde{S}_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}B'$.

Proof

From (3.29)

$$n\hat{\Sigma}_{11} = A_1'MA_1$$

We wish to apply Lemma 2.9 with $S = n\hat{\Sigma}_{11}$, $A = A_1$, $M = M$, $\Sigma = \Sigma_{1.2}$

Now $M^2 = M_1 + M_2^2$ from (3.32)

Hence $tr(M^2) = tr(M_1) + tr(M_2^2)$

$= n - p_2 - 1 + tr(M_2^2)$ from (3.20)

$= n - p_2 - 1 + tr(\tilde{S}_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}\tilde{S}_{22s}^{-1})$ c.f. (3.21) (3.33)

Also $\bar{A}_1'M^2\bar{A}_1 = \bar{A}_1'P_WM^2P_W\bar{A}_1$

$= BA_2'P_WM^2P_WA_2B'$ from (2.36)

$= BA_2'M^2A_2B'$

$= BA_2'M_1A_2B' + BA_2'M_2^2A_2B'$ from (3.32)

$= BA_2'M_2^2A_2B'$ from (3.24)

$= n^2B\tilde{S}_{22}(A_2'P_WA_2)^{-1}\tilde{S}_{22}B'$ from (3.19)

$= nBS_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}B' = n\psi$ (3.34)

Substituting (3.33) and (3.34) into Lemma 2.9 and dividing by n^2 gives the required result.

We might now obtain the moments of $\hat{\mu}_1$ and $\hat{\Sigma}_{11}$ conditional just on $\underline{x}_2$ and unconditionally as in Chapter 2. However, this does not offer much intuitive clarification and we do not propose to do so. Having approximately removed the bias conditional on s and $\underline{x}_2$ the bias conditional on $\underline{x}_2$ or unconditionally will also be approximately zero. We note that the unconditional moments of $\hat{\mu}_1$ and $\hat{\Sigma}_{11}$ under srs are given by Morrison (1971) for the case $p_1 = 1$. Birnbaum et al (1950) gives the unconditional variance of $\hat{\mu}_1$ for the case $p_1 = p_2 = 1$, $N = \infty$.

3.3 Model-Based Prediction

We now consider the optimal prediction of $\bar{x}_1$ and S_{11} under Model I conditional on s and $\underline{x}_2$. By 'optimal' we shall mean minimum variance unbiased within a given class of predictors.

Definition 3.1 : If X and Y are $p \times 1$ random vectors then we write $V(X) \leq V(U)$ iff $V(U) - V(X)$ is non-negative definite.

Definition 3.2 : Let $U(T)$ be the class of unbiased predictors of $T = T(\underline{x}_1)$, i.e.

$$U(T) = \{\tilde{T}(\underline{x}_{1s}) : E_I(\tilde{T}(\underline{x}_{1s}) - T(\underline{x}_1)|\underline{x}_2, s) = 0\}$$

$\hat{T} = \hat{T}(\underline{x}_{1s})$ is said to be a *minimum variance unbiased predictor* of T within a subclass $U^*(T)$ of $U(T)$ if

(i) $T \in U^*(T)$

and

(ii) $V_I(\hat{T}-T|\underline{x}_2, s) \leq V_I(\tilde{T}-T|\underline{x}_2, s)$ for all $\tilde{T} \in U^*(T)$

Two possible approaches to obtaining minimum variance unbiased predictors of $\bar{x}_1$ and S_{11s} would be:

(1) obtain a minimum variance predictor from a restricted class of unbiased predictors e.g. linear predictors for $\bar{x}_1$ or quadratic predictors for S_{11} by standard Lagrange multiplier techniques;

(2) obtain a minimum variance predictor from the while class of unbiased predictors by making restrictions on the distributional forms in Model I and by using a Lehmann -Scheffé type argument.

Approach (1) is applied to the prediction of $\bar{x}_1$, by a straightforward extension of linear prediction theory for univariate means, in Theorem 3.6. This approach would, however, be extremely laborious for the prediction of S_{11s}. Even for the simple case of $p_1 = 1$ with no mean structure for $X_1|X_2$, Mukhopadhyay (1978) resorts to making normal distributional assumptions. Having given Theorem 3.6 we shall therefore adopt Approach (2).

Theorem 3.6

$\hat{\mu}_1$ has minimum variance amongst the class of linear unbiased predictors of $\bar{x}_1$.

Proof.

Let a be an arbitrary $p_1 \times 1$ vector of constants. Then $a'\bar{x}_1$ is the (univariate) mean of the variate $a'x_{1i}$. Hence from e.g. Royall (1976, Theorem 2.1) the minimum variance linear unbiased predictor of $a'\bar{x}_1$ is

$$fa'\bar{x}_{1s} + \sum_{n+1}^{N} (a'\hat{\mu}_{1.2} + a'\hat{B}x_{2i})/N$$

where $\hat{\mu}_{1.2}$ and $\hat{B}$ are given in (3.3) and (3.4) and $f = n/N$.

$$= a'[f\bar{x}_{1s} + (1-f)(\bar{x}_{1s}-\hat{B}x_{2s}) + \hat{B}(\bar{x}_2-f\bar{x}_{2s})]$$

$$= a'(\bar{x}_{1s}-\hat{B}(\bar{x}_{2s}-\bar{x}_2))$$

$$= a'\hat{\mu}_1$$

Let $\tilde{\bar{x}}_1$ be any linear unbiased predictor of $\bar{x}_1$. Then $a'\tilde{\bar{x}}_1$ is a linear unbiased predictor of $a'\bar{x}_1$ and so

$$V_I(a'\tilde{\bar{x}}_1-a'\bar{x}_1|s,\underline{x}_2) \geq V_I(a'\hat{\mu}_1-a'\bar{x}_1|s,\underline{x}_2)$$

$$\therefore \quad a'V_I(\tilde{\bar{x}}_1-\bar{x}_1|s,\underline{x}_2)a \geq a'V_I(\hat{\mu}_1-\bar{x}_1|s,\underline{x}_2)a$$

But a is arbitrary and $\hat{\mu}_1$ is a linear unbiased predictor of $\bar{x}_1$. Hence from Definitions 3.1 and 3.2 $\hat{\mu}_1$ is a minimum variance linear unbiased predictor of $\bar{x}_1$.

We now adopt approach (2). We shall make the assumption that $X_I|X_2$ is normal in Model I and then apply the following result which is an extension of the Lehman Scheffé Theorem (e.g. Mood et al., 1974,p.326).

Lemma 3.7

Let Y be an observed random vector with distribution indexed by θ. Let $S = S(Y)$ be a complete sufficient statistic for θ. Let Z be an unobserved random vector such that the joint distribution of (Y,Z) is also indexed by θ. Suppose S is predictive sufficient for Z (see definition 1.3). Then if $\hat{T} = \hat{T}(S)$ is an unbiased predictor of $T = T(S,Z)$, $\hat{T}$ is the unique minimum variance unbiased predictor of T.

Proof

Let $\tilde{T} = \tilde{T}(Y)$ be any unbiased predictor of T. Then

$$V(\tilde{T}-T) = V\left[E(\tilde{T}-T|S,Z)\right] + E\left[V(\tilde{T}-T|S,Z)\right]$$

$$= V\left[E(\tilde{T}|S) - T\right] + E\left[V(\tilde{T}-T|S,Z)\right]$$

since S is predictive sufficient for Z and T is a given function of S and Z.

Now $E\left[V(\tilde{T}-T|S,Z)\right]$ is non-negative definite since $V(\tilde{T}-T|S,Z)$ is non-negative definite. Hence

$$V(\tilde{T}-T) \geq V\left[E(\tilde{T}|S) - T\right] \qquad (3.35)$$

But $$E\left[E(\tilde{T}|S)\right] = E(\tilde{T}) = E(T)$$

and $$E(\hat{T}) = E(T)$$

Hence $$E\left[E(\tilde{T}|S) - \hat{T}\right] = 0$$

But $E(\tilde{T}|S)$ and $\hat{T}$ are both functions of S, a complete sufficient statistic for θ and so $\hat{T} = E(\tilde{T}|S)$. The result follows from (3.35).

Note that Lemma 3.7 only applies to statistics, T, which are functions of S and Z. If we substitute $T(Y,Z)$ for $T(S,Z)$ the lemma would be invalid.

Counter example: Let $Y = (Y_1 \ldots Y_n)'$ where $Y_i \sim NID(\mu, 1)$ then $S = \bar{Y}$ is a complete sufficient statistic for μ, but Y_1, the minimum variance unbiased predictor of $T(Y) = Y_1$, is not a function of S.

Lemma 3.7 is sufficient for our purposes but we note that if we did wish to predict a statistic, $T(Y,z)$, which was not a function of S and z, we would attempt to write

$$T(Y,z) = U(Y) + V(S,z) \qquad (3.36)$$

and then set $\hat{T}(Y,z) = U(Y) + \hat{V}(S,z)$. Sometimes this is not possible.

Example (Mukhopadhyay, 1978) : Let $Y = (Y_1 \ldots Y_n)$, $z = (Y_{n+1} \ldots Y_N)$ where $Y_i \sim NID(0, \sigma^2)$. Then $S = \sum_1^n Y_i^2$ is a complete sufficient statistic for σ^2 but $T = \sum_1^N (Y_i - \bar{Y})^2/N$ cannot be written as (3.3.6).

We suggest that such an example is very contrived. If we were confident in setting $E(Y_i) = 0$ in the model then we would presumably be more interested in predicting $\Sigma Y_i^2/N$ than in taking the sum of square d deviations about $\bar{Y}$.

We now apply Lemma 3.7 to the prediction of $\bar{x}_1$ and S_{11}.

Theorem 3.8

If X_1 has a multivariate normal distribution given X_2 in Model I then the minimum variance unbiased predictors of $\bar{x}_1$ and S_{11} are:

$$\hat{\bar{x}}_1 = \hat{\mu}_1 \qquad \text{(defined in 3.3)}$$

$$\hat{S}_{11} = \hat{\hat{\Sigma}}_{11} \qquad \text{(defined in 3.27)}$$

Proof

If $X_1|X_2$ is normal then the distribution of $\underline{x}_{1s}$ given s and $\underline{x}_2$ is indexed by $\theta = (\mu_{1.2}, B, \Sigma_{1.2})$. From standard theory for the multivariate linear model (e.g. Arnold, 1981, Ch.19) a complete sufficient statistic for θ is given by $\hat{\theta} = (\hat{\mu}_{1.2}, \hat{B}, \hat{\Sigma}_{1.2})$. In Lemma 3.7 set $Y = \underline{x}_{1s}$, $Z = \underline{x}_{1s}^-$, $S = \hat{\theta}$. S is trivially predictive sufficient

for Z because $\underline{x}_{1s}$ and $\underline{x}_{1\bar{s}}$ are independent. Hence the conditions of the Lemma apply.

Firstly let $T = \bar{x}_1$. Then

$$T = f\bar{x}_{1s} + \sum_{i\bar{\in}s} x_{1i}/N$$

$$= n(\hat{\mu}_{1.2} + \hat{B}\bar{x}_{2s}) + \sum_{i\bar{\in}s} x_{1i}/N$$

Hence T is a function of S and Z (since $\underline{x}_2$ is considered fixed and known). Further $\hat{\mu}_1$ is a function of S since

$$\hat{\mu}_1 = \hat{\mu}_{1.2} + \hat{B}\bar{x}_2$$

and from Theorem 3.3

$$E_I(\hat{\mu}_1 | s, \underline{x}_2) = \mu_1 + B(\bar{x}_2 - \mu_2)$$

Setting $n = N$ in Theorem 2.1 we obtain

$$E_I(\bar{x}_1 | s, \underline{x}_2) = \mu_1 + B(\bar{x}_2 - \mu_2)$$

Hence $\hat{\mu}_1$ is an unbiased predictor of $\bar{x}_1$ and from Lemma 3.7. $\hat{\mu}_1$ is the minimum variance unbiased predictor of $\bar{x}_1$.

Now let $T = S_{11}$

$$(N-1)S_{11} = \sum_1^N x_{1i}x'_{1i} - N\bar{x}_1\bar{x}'_1$$

$$= \sum_1^n (x_{1i}-\bar{x}_{1s})(x_{1i}-\bar{x}_{1s})' + n\,\bar{x}_{1s}\bar{x}'_{1s} + \sum_{n+1}^N x_{1i}x'_{1i} - N\bar{x}_1\bar{x}'_1$$

$$= n(\hat{\Sigma}_{1.2} + \hat{B}\,\tilde{S}_{22s}\,\hat{B}') + n\,\bar{x}_{1s}\,\bar{x}'_{1s} + \sum_{n+1}^N x_{1i}\,x'_{1i} - N\,\bar{x}_1\,\bar{x}'_1$$

We have already noted that $\bar{x}_{1s}$ and $\bar{x}_1$ are functions of $S = \hat{\theta}$ and $Z = \underline{x}_{1\bar{s}}$. Hence S_{11} is also a function of S and Z . Also $\hat{S}_{11} = \hat{\Sigma}_{11}$ is a function of S since

$$\hat{\hat{\Sigma}}_{11} = \lambda\hat{\Sigma}_{1.2} + \hat{B}\, S_{22}\, \hat{B}'$$

where λ is a known constant

From (3.28)

$$E_I(\hat{S}_{11}|s, \underline{x}_2) = \Sigma_{11} + B(S_{22} - \Sigma_{22})B'$$

Setting $n = N$ in Theorem 2.3 we obtain

$$E_I(S_{11}|s, \underline{x}_2) = \Sigma_{11} + B(S_{22} - \Sigma_{22})B'$$

Hence $\hat{S}_{11}$ is an unbiased predictor of S_{11} and, from Lemma 3.7, $\hat{S}_{11}$ is the minimum variance unbiased predictor of S_{11}.

Note that if the design is balanced on $\bar{x}_{2s}$, i.e. $\bar{x}_{2s} = \bar{x}_2$ then

$$\hat{\bar{x}}_1 = \hat{\mu}_1 = \bar{x}_{1s}$$

and if the design is balanced on S_{22s}, i.e. $S_{22s} = S_{22}$ then $tr(S_{22}\, S_{22s}^{-1}) = p_2$, $\lambda = n/(n-1)$ and

$$\hat{S}_{11} = n\hat{\Sigma}_{11}/(n-1) = S_{11s}$$

Hence the 'naive' predictors, $\bar{x}_{1s}$ and S_{11s}, considered in Chapter 2 are optimal in these special cases.

It is desirable, for purposes of subsequent analysis, that $\hat{\Sigma}_{11}$ and $\hat{S}_{11}$ be non-negative definite.

Lemma 3.9

$\hat{\Sigma}_{11}$ is non-negative definite

$\hat{S}_{11}$ is non-negative definite provided $tr(S_{22}\, S_{22s}^{-1}) \leq n - 1$

Proof

$\hat{\Sigma}_{1.2}$ and $\hat{B}\, S_{22}\, \hat{B}'$ are always non-negative definite and hence so is $\hat{\Sigma}_{11}$. From (3.27) $\hat{S}_{11}$ is non-negative definite if $\lambda \geq 0$ i.e. if $tr(S_{22}\, S_{22s}^{-1}) \leq n - 1$.

The condition $tr(S_{22}\ S_{22s}^{-1}) \lesssim n - 1$ should hold except for very small or extreme samples. If this condition did not hold it would seem dangerous to attempt to predict S_{11} at all.

We now obtain the first and second moments of $\hat{\bar{x}}_1$ and $\hat{S}_{11}$ as predictors of $\bar{x}_1$ and S_{11} .

Theorem 3.10

$$E_I(\hat{\bar{x}}_1 - \bar{x}_1 | s, \underline{x}_2) = 0$$

$$V_I(\hat{\bar{x}}_1 - \bar{x}_1 | s, \underline{x}_2) = [1 - f + (\bar{x}_{2s}-\bar{x}_2)' \tilde{S}_{22}^{-1} (\bar{x}_{2s}-\bar{x}_2)]\Sigma_{1.2}/n$$

Proof

We have already shown that $\hat{\bar{x}}_1$ is prediction unbiased for $\bar{x}_1$ in Theorem 3.8. To obtain the prediction variance we write as in (3.13).

$$\hat{\bar{x}}_1 - \bar{x}_1 = \sum_1^N w_i x_{1i}$$

where
$$w_i = [1 - (x_{2i}-\bar{x}_{2s})' \tilde{S}_{22s}^{-1}(\bar{x}_{2s}-\bar{x}_2)]/n - \frac{1}{N} \qquad i = 1...n$$

$$= - 1/N \qquad i = n+1...N$$

$$\therefore \quad V_I(\hat{\bar{x}}_1-\bar{x}_1 | s, \underline{x}_2) = \Sigma w_i^2 \Sigma_{1.2}$$

$$= \Big[n(1-f)^2/n^2 + \sum_s [(x_{2i}-\bar{x}_{2s})' \tilde{S}_{22s}^{-1}(x_{2s}-\bar{x}_2)]^2 /n^2 + (N-n)/N^2\Big]\Sigma_{1.2}$$

$$= [1-f+(\bar{x}_{2s}-\bar{x}_2)' \tilde{S}_{22s}^{-1}(\bar{x}_{2s}-\bar{x}_2)]\Sigma_{1.2}/n$$

as required

Comparing this result with Corollary 2.2 we see that the prediction bias of $\bar{x}_{1s}$ has been reduced to zero in $\hat{\bar{x}}_1$. The variance has been increased due to the estimation of B. The amount of increase depends on both $\bar{x}_{2s} - \bar{x}_2$ (which is related to the difference between $\hat{\bar{x}}_1$ and $\bar{x}_{1s}$) and S_{22s}^{-1} the usual factor in the OLS estimate of B .

Theorem 3.11

$$E_I(\hat{S}_{11}-S_{11}|s,\underline{x}_2) = 0$$

If $X_1|X_2$ is multivariate normal then

$$COV_I\{(\hat{S}_{11}-S_{11})_{ij}, (\hat{S}_{11}-S_{11})_{k\ell}|s,\underline{x}_2\} = \gamma[\Sigma_{1.2ik}\Sigma_{1.2j\ell} + \Sigma_{1.2i\ell}\Sigma_{1.2jk}]/(n-1)$$
$$+[\Sigma_{1.2j\ell}\psi_{ik} + \Sigma_{1.2i\ell}\psi_{jk} + \Sigma_{1.2jk}\psi_{i\ell} + \Sigma_{1.2ik}\psi_{j\ell}]/(n-1)$$

where

$$\gamma = (n-p_2-1)(n-1)\lambda^2/n^2 + tr(S_{22}S_{22s}^{-1}\tilde{S}_{22}S_{22s}^{-1})/(n-1) - (n-1)/(N-1)$$

$$\psi = BS_{22}S_{22s}^{-1}S_{22}B' - (n-1)BS_{22}B'/(N-1)$$

Proof

We have already shown that $\hat{S}_{11}$ is prediction-unbiased for S_{11} in Theorem 3.8. By analogy with (2.30)-(2.32) let

$$A_{1N}' = (x_{11}\cdots x_{1N})$$

$$P_{NN} = I_N - 1_N 1_N'/N$$

Then
$$(N-1)S_{11} = A_{1N}'P_{NN}A_{1N} \qquad (3.37)$$

From (3.27)

$$\hat{S}_{11} = \hat{\tilde{\Sigma}}_{11} = \lambda\,\hat{\Sigma}_{1.2} + \hat{B}S_{22}\hat{B}'$$

$$= \lambda\, A_1'M_1A_1/n + NA_1'M_2A_1/(N-1)n$$

from (3.16) and (3.17)

$$= A_1'HA_1 \quad \text{say} \qquad (3.38)$$

where
$$H = \alpha M_1 + \beta M_2 \qquad (3.39)$$

$$\alpha = \lambda/n\ ,\ \beta = N/(N-1)n$$

Let H^* be the $N \times N$ matrix which has H at the top left $n \times n$ corner and zeros elsewhere

$$H^* = \begin{pmatrix} H & 0 \\ 0 & 0 \end{pmatrix}$$

Then from (3.37) and (3.38)

$$\hat{S}_{11} - S_{11} = A_{1N}'(H^* - P_{NN}/(N-1))A_{1N}$$

$$= A_{1N}'QA_{1N} \text{ , say}$$

We now apply Lemma 2.9 with $S = \hat{S}_{11} - S_{11}$, $A = A_{1N}$, $M = Q$, $\Sigma = \Sigma_{1.2}$. Now

$$Q^2 = (H^*)^2 - H^* P_{NN}/(N-1) - P_{NN}H^*/(N-1) + P_{NN}/(N-1)^2 \qquad (3.40)$$

since P_{NN} is idempotent

In obvious notation

$$(H^*)^2 = (H^2)^* \qquad (3.41)$$

Partition P_{NN} into

$$P_{NN} = \begin{pmatrix} P_{nn} & P_{n\bar{n}} \\ P_{\bar{n}n} & P_{\bar{n}\bar{n}} \end{pmatrix}$$

where P_{nn} is $n \times n$, $P_{n\bar{n}}$ is $n \times (N-n)$ etc.

Then
$$H^*P_{NN} = \begin{pmatrix} HP_{nn} & HP_{n\bar{n}} \\ 0 & 0 \end{pmatrix} \qquad (3.42)$$

But from (3.18) and (3.19)

$$M_1 1_n = M_2 1_n = 0$$

Hence from (3.39)

$$H1_n = 0 \qquad (3.43)$$

and so

$$HP_{nn} = 0 \tag{3.44}$$

Also

$$P_{nn} = I_n - 1_n 1_n'/N$$

$$= P_W + 1_n 1_n'(N-n)/Nn$$

where P_W is defined in (2.32)

Hence from (3.43)

$$HP_{nn} = HP_W$$

but from (3.18) and (3.19)

$$M_1 P_W = M_1 \quad , \quad M_2 P_W = M_2$$

Hence from (3.39)

$$HP_{nn} = HP_W = H \tag{3.45}$$

Hence from (3.42), (3.44) and (3.45)

$$H^* P_{NN} = H^* \tag{3.46}$$

Similarly

$$P_{NN} H^* = H^* \tag{3.47}$$

Substituting (3.41), (3.46) and (3.47) into (3.40) gives

$$Q^2 = \left[H^2 - 2H/(N-n) \right]^* + P_{NN}/(N-1)^2 \tag{3.48}$$

Hence

$$tr(Q^2) = tr(H^2) - 2tr(H)/(N-1) + 1/(N-1) \tag{3.49}$$

Now $\quad tr(H) = \alpha tr(M_1) + \beta tr(M_2)$ from (3.39)

$= \alpha(n-p_2-1) + \beta tr(\tilde{S}_{22}\tilde{S}_{22s}^{-1})$ from (3.20) and (3.21)

$= 1 - tr(S_{22}\tilde{S}_{22s}^{-1})/n + tr(S_{22}\tilde{S}_{22s}^{-1})/n$

$= 1$ (3.50)

$tr(H^2) = \alpha^2 tr(M_1) + \beta^2 tr(M_2^2)$ from (3.31)

$= \alpha^2(n-p_2-1) + \beta^2 tr(\tilde{S}_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}\tilde{S}_{22s}^{-1})$ from (3.20) and (3.33)

$= (n-p_2-1)\lambda^2/n^2 + tr(S_{22}\tilde{S}_{22s}^{-1}S_{22}\tilde{S}_{22s}^{-1})/n^2$

$= (n-p_2-1)\lambda^2/n^2 + tr(S_{22}S_{22s}^{-1}S_{22}S_{22s}^{-1})/(n-1)^2$ (3.51)

Combining (3.49) - (3.51)

$tr(Q^2) = (n-p_2-1)\lambda^2/n^2 + tr(S_{22}S_{22s}^{-1}S_{22}S_{22s}^{-1})/(n-1)^2 - 1/(N-1)$

$= \gamma/(n-1)$ (3.52)

Let $\quad \bar{A}_{1N} = E_I(A_{1N}|s,\underline{x}_2)$

Then $\quad \bar{A}_{1N}'Q^2\bar{A}_{1N} = \bar{A}_1'[H^2-2H/(N-1)]\bar{A}_1 + \bar{A}_{1N}'P_{NN}\bar{A}_{1N}/(N-1)^2$

from (3.48), where $\bar{A}_1$ is defined in (2.36).

Now $\bar{A}_{1N}'P_{NN}\bar{A}_{1N} = (N-1)BS_{22}B'$ as in the proof of Theorem 2.4

$A_1'HA_1 = \alpha A_1'M_1A_1 + \beta A_1'M_2A_1$ from (3.39)

$= \alpha BA_2'M_1A_2B' + \beta BA_2'M_2A_2B'$ from (2.36), (3.18) and (3.19)

$= n\beta B\tilde{S}_{22}B'$ from (3.24) and (3.25)

$\bar{A}_1'H^2\bar{A}_1 = \alpha^2\bar{A}_1'M_1\bar{A}_1 + \beta^2\bar{A}_1'M_2^2\bar{A}_1$ from (3.32)

$= n\beta^2B\tilde{S}_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}B'$ from (3.34)

Hence
$$\bar{A}'_{1N}Q^2\bar{A}_{1N} = n\beta^2 B\tilde{S}_{22}\tilde{S}^{-1}_{22s}\tilde{S}_{22}B' - 2n\beta B\tilde{S}_{22}B'/(N-1)$$
$$+ BS_{22}B'/(N-1)$$
$$= BS_{22}S^{-1}_{22s}S_{22}B'/(n-1) - BS_{22}B'/(N-1)$$
$$= \psi/(n-1) \qquad (3.53)$$

The result then follows by substituting (3.52) and (3.53) into Lemma 2.9.

Comparing this result with Corollary 2.5 we see that the prediction bias of S_{11s} has been reduced to zero in S_{11s}. From Theorem 2.10 the variances of both S_{11s} and $\hat{S}_{11}$ are of $O(n^{-1})$ and so the MSE of S_{11s} is of $O(1)$ compared with the MSE of $\hat{S}_{11}$ which is of $O(n^{-1})$.

Note finally that if $X_1|X_2$ is normal then $\hat{\bar{x}}_1 - \bar{x}_1$ will be normally distributed given s and $\underline{x}_2$ with mean vector and covariance matrix given in Theorem 3.10. As for $\hat{\Sigma}_{11}$ the distribution of $\hat{S}_{11}$ will be complicated.

3.4 Design-based Estimation

In this section we consider some of the estimators of $\bar{x}_1$ and S_{11} suggested in the literature on the basis of their properties with respect to the randomisation distribution induced by $p(s|\underline{x}_2)$. We begin with a general sampling design, $p(s|\underline{x}_2)$, and then consider the special cases of srs and stratified srswor. We shall not consider estimators based on a combination of model-based and randomisation-based arguments (e.g. Liu, 1974a; Chaudhuri, 1978; Zacks and Solomon, 1981; Zacks, 1981).

(i) The most common design-unbiased estimator of $\bar{x}_1$ is (Horvitz and Thompson, 1952):

$$e_1(\bar{x}_1) = \sum_s x_{1i}/N\pi_i$$

where
$$\pi_i = \pi_i(\underline{x}_2) = \sum_{s \ni i} p(s|\underline{x}_2)$$

assuming
$$\pi_i > 0 \quad i = 1...N$$

The choice of a corresponding estimator of S_{11} is not obvious. Liu(1974a, b) proposes the design-unbiased estimator:

$$e_2(S_{11}) = \sum_s x_{1i}x_{1i}'/N\pi_i - \sum_{i\neq j\in s} x_{1i}x_{ij}'/\pi_{ij}N(N-1)$$

where $$\pi_{ij} = \pi_{ij}(\underline{x}_2) = \sum_{s\ni i,j} p(s|\underline{x}_2)$$

assuming $$\pi_{ij} > 0 \quad i,j = 1...N$$

Chaudhuri (1978) notes that $e_2(S_{11})$ may be negative (in the univariate case) and following Murthy (1963), proposes:

$$e_3(S_{11}) = \sum_{i<j\in s} (x_{1i}-x_{1j})(x_{1i}-x_{1j})'/\pi_{ij}N(N-1)$$

which is non-negative definite unbiased

(ii) Alternative design-unbiased estimators are also given by Chaudhuri (1978):

Let Let $h_i(s) = 1$ if $i \in s$

$= 0$ if $i \notin s$

Let $$t_i = \sum_s h_i(s) \quad , \quad t_{ij} = \sum_s h_i(s)h_j(s)$$

then unbiased estimators of $\bar{x}_1$ and S_{11} are

$$e_4(\bar{x}_1) = \sum_s x_{1i}/Nt_ip(s|\underline{x}_2)$$

$$e_5(S_{11}) = \sum_{i<j\in s} (x_{1i}-x_{ij})(x_{1i}-x_{ij})'/N(N-1)t_{ij}\ p(s|\underline{x}_2)$$

assuming $t_i, t_{ij} > 0$ (note $t_i \geq \pi_i$, $t_{ij} \geq \pi_{ij}$) . Special cases of these estimators are also given by Murthy (1963).

(iii) If the sampling design is *with replacement*, say we make n independent selections where the i^{th} element of the population is chosen with probability p_i at each selection, then unbiased estimators of $\bar{x}_1$ and S_{11} are:

$$e_6(\bar{x}_1) = \sum_s x_{1i}/nNp_i \qquad \text{(Hansen and Hurvitz, 1943)}$$

$$e_7(S_{11}) = \left[\sum_s x_{1i}x'_{1i}/p_i - \sum_{i\neq j\in s} x_{1i}x'_{1i}/p_ip_jN(n-1)\right]/n(N-1)$$

(Das and Tripathi, 1977)

or

$$e_8(S_{11}) = \left[\sum_s 1/p_i \sum_s x_{1i}x'_{1i}/p_i - \sum_s x_{1i}/p_i \sum_s x'_{1i}/p_i\right]/n(n-1)\ N(N-1)$$

(Rao, 1975)

(iv) If N is large we may approximate π_i by np_i and $e_6(\bar{x}_1)$ becomes equal to $e_1(\bar{x}_1)$, $e_7(S_{11})$ becomes equal to $e_2(S_{11})$ (if $n(n-1)p_ip_j$ is also replaced by π_{ij}) and $e_8(S_{11})$becomes

$$e_9(S_{11}) = n\left[\sum_s 1/\pi_i \sum_s x_{1i}x'_{1i}/\pi_i - \sum_s x_{1i}/\pi_i \sum_s x'_{1i}/\pi_i\right]/(n-1)N^2$$

(v) Some asymptotically design-unbiased estimators may be preferable to exactly unbiased estimators. For example, Särndal (1980) argues that the 'consistent ratios estimator' of Brewer (1963) and Hájek (1971):

$$e_{10}(\bar{x}_1) = \left(\sum_s x_{1i}/\pi_i\right)/\left(\sum_s 1/\pi_i\right)$$

is preferable to $e_1(\bar{x}_1)$. Similarly, by analogy with $e_9(S_{11})$ we might consider

$$e_{11}(S_{11}) = \sum_s x_{1i}x'_{1i}/N\pi_i - \left(\sum_s x_{1i}/N\pi_i \sum_s x'_{1i}/N\pi_i\right)/\left(\sum_s 1/N\pi_i\right)$$

(Nathan and Holt, 1980, 3.2).

or by analogy with $e_3(S_{11})$

$$e_{12}(S_{11}) = \left[\sum_{i<j\in s} (x_{1i}-x_{1j})(x_{1i}-x_{1j})'/\pi_{ij}\right]/\left[\sum_{i\neq j\in s} 1/\pi_{ij}\right]$$

(vi) When auxiliary information $\underline{x}_2$ is available and $p_2 = 1$ Das and Tripathi (1978) propose multiplicative adjustments to given estimators, e, of the form

$$e^* = e\cdot(\bar{x}_2/\bar{x}_{2s})^{\alpha}$$

$$e^{**} = e\cdot(S_{22}/S_{22s})^{\alpha}$$

$$e^{***} = e\cdot(S_{22}\bar{x}_{2s}^2/S_{22s}\bar{x}_2^2)^{\alpha}$$

where α is a chosen constant. Similarly Nathan and Holt (1980) suggest using probability weights in the MLE's of Section 3.2.

$$\hat{\mu}_1^* = e(\bar{x}_1) - \hat{B}\left(e(\bar{x}_2) - \bar{x}_2\right)$$

$$\hat{\Sigma}_{11} = e(S_{11}) - \hat{B}\left(e(S_{22}) - S_{22}\right)\hat{B}'$$

$\hat{B}$ may also be a probability-weighted estimator of B .

Example 1 : srswor

In this case

$e_1(\bar{x}_1) = e_4(\bar{x}_1) = e_{10}(\bar{x}_1) = \bar{x}_{1s}$

$e_2(S_{11}) = e_3(S_{11}) = e_5(S_{11}) = e_9(S_{11}) = ne_{11}(S_{11})/(n-1) = e_{12}(S_{11}) = S_{11s}$

We might use srswor if we believed Model I to be

$$E_I(X_{1i}|X_{2i}) = \mu \qquad V_I(X_{1i}|X_{2i}) = \Sigma$$

In this case X_2 is just a scalar constant and so $\bar{x}_{2s} - \bar{x}_2 = 0$, $S_{22s} = S_{22} = 0$, $\lambda = n/(n-1)$ and the optimal model-based predictors of Theorem 3.8 are just $\bar{x}_{1s}$ and S_{11s} as above.

Example 2 : Stratified srswor

With the same notation as in Section 2.2.2.

$$e_1(\bar{x}_1) = e_4(\bar{x}_1) = e_{10}(\bar{x}_1) = \Sigma W_h\ \bar{x}_{1h} = \tilde{\bar{x}}_1 \text{ , say}$$

where

$$\bar{x}_{1h} = \sum_{i\in s\cap S_h} x_{1i}/n_h$$

This is the usual design-unbiased estimator of $\bar{x}_1$ for stratified sampling.

$$e_2(S_{11}) = e_3(S_{11}) = e_5(S_{11}) = e_{12}(S_{11})$$
$$= N\left\{\Sigma\alpha_h s_h + \Sigma W_h(\bar{x}_{1h} - \tilde{\bar{x}}_1)(\bar{x}_{1h} - \tilde{\bar{x}}_1)'\right\}/(N-1) \qquad (3.54)$$

where

$$s_h = \sum_{i\epsilon s\cap S_h} (x_{1i} - \bar{x}_{1h})(x_{1i} - \bar{x}_{1h})'/n_h$$

$$\alpha_h = W_h + W_h(N_h - n_h)/N(n_h - 1)$$

This estimator has also been proposed specifically for the case of stratified srswor by Koop (1970) and Gupta et al (1979).

$$(n-1)\, e_9(S_{11})/n = e_{11}(S_{11}) = \Sigma W_h s_h + \Sigma W_h(\bar{x}_{1h} - \tilde{\bar{x}}_1)(\bar{x}_{1h} - \tilde{\bar{x}}_1)'$$

Wakimoto (1971 a, b) has proposed a similar estimator for stratified srsw r :

$$\sum_h\left(W_h + W_h^2/(n_h-1)\right)s_h + \Sigma W_h(\bar{x}_{1h} - \tilde{\bar{x}}_1)(\bar{x}_{1h} - \tilde{\bar{x}}_1)'$$

This estimator was also proposed by Aoyama (1954) for the case of proportional allocation with $H = 2$.

In the case of a stratified population we might adopt the model described in Section 2.2.2. If we make the additional assumption of normality we might write this model as:

$$X_1|X_2 = e_h \sim N_{p_1}(\mu_{1.2} + B_h,\ \Sigma_{1.2})$$

Substituting into (2.14), (2.16), (3.27) and (3.4) we obtain

$$\bar{x}_{1s} = \Sigma w_h \bar{x}_{1h}$$

$$\hat{B}_h = \bar{x}_{1h} - \bar{x}_{11}$$

$$S_{11s} = n\left[\Sigma w_h S_h + \Sigma w_h(\bar{x}_{1h} - \bar{x}_{1s})(\bar{x}_{1h} - \bar{x}_{1s})'\right]/(n-1)$$

$$\lambda = n\left[n(N-1) - N\Sigma W_h(1-W_h)/w_h\right]/(n-1)(N-1)(N-H-2)$$

Hence from Theorem 3.8 optimal predictors of $\bar{x}_1$ and S_{11} under this model are

$$\hat{\bar{x}}_1 = \Sigma W_h \bar{x}_{1h}$$

$$\hat{S}_{11} = (n-1)\lambda\Sigma w_h s_h/n + N\Sigma W_h(\bar{x}_{1h} - \hat{\bar{x}}_1)(\bar{x}_{1h} - \hat{\bar{x}}_1)'$$

Now $\hat{\bar{x}}_1 = \tilde{\bar{x}}_1$ but $\hat{S}_{11}$ differs from the design-based predictors of S_{11} above by applying a weight w_h to s_h. If, however, we modify the model to

$$X_1|X_2 = e_h \sim N_{p_1}(\mu_{1.2} + B_h, \Sigma_h)$$

to allow for different within stratum covariance matrices, which would usually be more acceptable to most survey samplers, we may again obtain optimal predictors of $\bar{x}_1$ and S_{11} using Lemma 3.7. For this model a complete sufficient statistic for $(\mu_{1.2}, B_2 \ldots B_H, \Sigma_1 \ldots \Sigma_H)$ is $(\bar{x}_{11} \ldots \bar{x}_{1H}\ s_1 \ldots s_H)$ and applying Lemma 3.7, minimum variance unbiased predictors of $\bar{x}_1$ and S_{11} are given by $\tilde{\bar{x}}_1$ and (3.54) respectively. This is another example of the 'duality' between finite population prediction theory and without replacement random sampling theory.

Returning to the general case, we do not wish to dwell on a comparison of the above estimators, but we do note the implications of a *location shift* $LS(K):X_1 \to X_1 + K$. We might expect that for estimators $e(\bar{x}_1)$ of $\bar{x}_1$ and $e(S_{11})$ of S_{11} :

$$LS(K) : e(\bar{x}_1) \to e(\bar{x}_1) + K \qquad (3.55)$$

$$LS(K) : e(S_{11}) \to e(S_{11}) \qquad (3.56)$$

In fact only $e_{10}(\bar{x}_1)$ obeys (3.55) in general (whereas e_1, e_4 and e_6 only do so in special cases) and only $e_3(S_{11})$, $e_5(S_{11})$, $e_8(S_{11})$, $e_9(S_{11})$, $e_{11}(S_{11})$ and $e_{12}(S_{11})$ obey (3.5.6) (whereas $e_2(S_{11})$ and $e_7(S_{11})$ do not in general).

Recalling the definition of A_1 in (2.30) we may write any of the above estimators of $\bar{x}_1$ as

$$e_{\underline{x}_1}(A_1) = A_1'h$$

where h is a $n \times 1$ vector of constants possibly depending on s, and any of the estimators of S_{11} as

$$e_{S_{11}}(A_1) = A_1'MA_1$$

where M is a $n \times n$ symmetric matrix of constants. Conditions (3.55) and (3.56) are then equivalent to

(1) $\quad h'1_n = 1$

(2) $\quad M1_n = 0$

If (1) holds

$$E_I(e_{\underline{x}_1}(A_1)|s, \underline{x}_2) = \mu_1 + B(e_{\underline{x}_1}(A_2) - \mu_2)$$

where A_2 is defined in (2.31).

This is a simple extension of the result for $e_{\underline{x}_1}(A_1) = \bar{x}_{1s}$ in Theorem 2.1.

If (2) holds

$$E_I(e_{S_{11}}(A_1)|s, \underline{x}_2) = Be_{S_{11}}(A_2)B' + tr(M)\Sigma_{1.2}$$

This would correspond to Theorem 2.4 for $e_{S_{11}}(A_1) = S_{11s}$ if $tr(M) = 1$. The only estimator of S_{11} obeying (2) for which $tr(M) = 1$ is $e_{12}(S_{11})$. We shall use this estimator in Chapter 4 and so give it special notation

$$S^*_{11s} = e_{12}(S_{11}) \qquad (3.57)$$

Note that, although this is essentially an arbitrary choice, S^*_{11s} is equal to most of the other design-based estimators in the above examples and is likely to be approximately equal to the other estimators for most designs. We record the moments of S^*_{11s} under Model I for use in Chapter 4.

Theorem 3.12

$$E_I(S^*_{11s}|s,\underline{x}_2) = \Sigma_{11} + B(S^*_{22s} - \Sigma_{22})B' \qquad (3.58)$$

where $S^*_{22s} = \left[\sum_{i<j\in s}(x_{2i}-x_{2j})(x_{2i}-x_{2j})'/\pi_{ij}\right] / \left[\sum_{i\neq j\in s} 1/\pi_{ij}\right]$

If $X_1|X_2$ is multivariate normal then

$$\begin{aligned}\mathrm{cov}_I(S^*_{11sij}, S^*_{11sk\ell}|s,\underline{x}_2) = \mathrm{tr}(M^2)(&\Sigma_{1.2ik}\Sigma_{1.2j\ell} + \Sigma_{1.2i\ell}\Sigma_{1.2jk} \\ &+ \Sigma_{1.2j\ell}\psi_{ik} + \Sigma_{1.2i\ell}\psi_{jk} + \Sigma_{1.2jk}\psi_{i\ell} + \Sigma_{1.2ik}\psi_{j\ell})\end{aligned} \qquad (3.59)$$

where

$$M_{ii} = \left[\sum_{\alpha\neq i} 1/\pi_{i\alpha}\right] / \left[\sum_{\alpha\neq\beta} 1/\pi_{\alpha\beta}\right]$$

$$M_{ij} = -1/\pi_{ij} \sum_{\alpha\neq\beta} 1/\pi_{\alpha\beta} \qquad i\neq j$$

$$\psi = B\, S^{**}_{22s}\, B'$$

$$S^{**}_{22s} = \left[\sum_{i<j\in s}(x_{2i}-x_{2j})(x_{2i}-x_{2j})'/h_{ij}\right] / \left[\sum_{i\neq j\in s} 1/h_{ij}\right]$$

$$h_{ij} = 1/(M^2)_{ij}$$

Proof

$$S^*_{11s} = \left[\sum_{i<j}(x_{1i}-x_{ij})(x_{1i}-x_{ij})'/\pi_{ij}\right] / \left[\sum_{\alpha\neq\beta} 1/\pi_{\alpha\beta}\right]$$

$$= \left[\sum_i x_{1i}x'_{1i} \sum_{\alpha\neq i} 1/\pi_{i\alpha} - \sum_{i\neq j} x_{1i}x'_{1j}/\pi_{ij}\right] / \left[\sum_{\alpha\neq\beta} 1/\pi_{\alpha\beta}\right]$$

$$= A_1'MA_1$$

(3.58) then follows from Lemma 2.3 since $M1_n = 0$, $\mathrm{tr}(M) = 1$ and

$$A_2'MA_2 = S^*_{22s}$$

(3.59) follows from Lemma 2.9 with

$$\psi = BA_2'M^2A_2B'/\mathrm{tr}(M^2)$$

Now

$$M1_n = 0 \implies M^2 1_n = 0 \implies$$

$$A_2' M^2 A_2 = - \sum_{i<j} (x_{2i}-x_{2j})(x_{2i}-x_{2j})' (M^2)_{ij}$$

$$= - S^{**}_{22s} \sum_{i \neq j} (M^2)_{ij}$$

Also $M^2 1_n = 0 \implies$

$$\sum_{i \neq j} (M^2)_{ij} = - \sum_{i} (M^2)_{ii} = - \mathrm{tr}(M^2)$$

Hence $\psi = B\, S^*_{22s}\, B'$ as required.

The conditional bias of S^*_{11s} given s and $\underline{x}_2$ is therefore non-zero in general (as for S_{11s}) although averaged over all possible samples s the bias, $E(S^*_{11s}|\underline{x}_2) - \Sigma_{11}$, is approximately zero (unlike S_{11s} in general).

3.5 Conclusion

In Chapter 2 we showed that the standard estimators $\bar{x}_{1s}$ and S_{11s} could be asymptotically biased for μ_1 and Σ_{11} (or $\bar{x}_1$ and S_{11}). In this chapter we have considered both design-based and model-based alternative estimators. We have seen in Theorem 3.12 that the design-based estimators can also be asymptotically biased, conditional on s and $\underline{x}_2$, although they will be approximately unbiased averaged over all possible samples s. The model-based estimators (predictors) on the other hand are asymptotically conditional unbiased provided the model is true.

We have also noted that in one important special case, stratified sampling, the design-based estimators are equal to the optimal model-based predictors under a simple model. Holt et al (1980b) gives an example where strata are determined by quantiles of a continuous univariate $\underline{x}_2$. They find that, although the design-based estimator is not as good as the estimator based on the true model, it is surprisingly efficient. We would argue that this is because the design-based estimator is an optimal estimator for a model which is not very far from the true model.

CHAPTER FOUR - MULTIVARIATE METHODS UNDER PEARSON-TYPE SELECTION SCHEME

In Chapters 2 and 3 we considered the estimation of μ_1 and Σ_{11} and the prediction of $\bar{x}_1$ and S_{11}. In this chapter we consider the estimation of functions of Σ_{11}, viz correlation coefficients (Section 4.1), regression coefficients (Section 4.2) and principal components (Section 4.3) and the estimation of parameters in a factor analysis model for Σ_{11} (in Section 4.4). We no longer consider the prediction problem,

4.1 Correlation Coefficients

In this section we consider the estimation of

$$P_{11} = \left[\Sigma_{11ij}/\sigma_{1i}\sigma_{ij}\right] \text{ , the correlation matrix of } X_1, \tag{4.1}$$

where

$$\sigma_{1i}^2 = \Sigma_{11ii} \tag{4.2}$$

and Σ_{11} is defined in (2.4).

We consider three estimators of P_{11}:

(1) the standard estimator:

$$R_{11} = \left[S_{11sij}/(S_{11sii}\ S_{11sjj})^{\frac{1}{2}}\right] \tag{4.3}$$

where S_{11s} is defined in (2.16)

(2) the MLE under joint normality of (X_1, X_2)

$$\hat{P}_{11} = \left[\hat{\Sigma}_{11ij}/(\hat{\Sigma}_{11ii}\ \hat{\Sigma}_{11jj})^{\frac{1}{2}}\right] \tag{4.4}$$

where $\hat{\Sigma}_{11}$ is defined in (3.2)

(3) a design-based estimator

$$R^*_{11} = \left[S^*_{11sij}/(S^*_{11sii}\ S^*_{11sjj})^{\frac{1}{2}}\right] \tag{4.5}$$

where S^*_{11s} is defined in (3.57).

We shall only consider the asymptotic bias of these estimators, which we referred to in Chapter 2 as the 'selection effect'.

The Standard Estimator

Theorem 4.1

If Model I holds and $V_I(S_{11s}|s, \underline{x}_2) = O(n^{-1})$ then

$$E_I(R_{11ij}|s,\underline{x}_2) = (P_{11ij}+\rho_i'\Delta\rho_j)(1+\rho_i'\Delta\rho_i)^{-\frac{1}{2}}(1+\rho_j'\Delta\rho_j)^{-\frac{1}{2}} + O(n^{-\frac{1}{2}}) \quad (4.6)$$

where

$$\rho_i = (\rho_{i1}....\rho_{ip_2})' \quad (4.7)$$

$$\rho_{ij} = \text{corr}_I\left[(X_1)_i, (X_2)_j\right]$$

$$= \Sigma_{12ij}/\sigma_{1i}\sigma_{2j} \quad (4.8)$$

$$\Delta = P_{22}^{-1}(D_2^{-1} S_{22s} D_2^{-1} - P_{22})P_{22}^{-1} \quad (4.9)$$

$$D_2 = \text{diag}(\sigma_{2i}) \quad (4.10)$$

$$\sigma_{2i}^2 = \Sigma_{22ii} \quad (4.11)$$

$P_{22} = D_2^{-1} \Sigma_{22} D_2^{-1}$, the correlation matrix of X_2 (4.12)

Proof

From Theorem 2.4

$$E_I(S_{11s}|s,\underline{x}_2) = \Sigma_{11} + B(S_{22s} - \Sigma_{22})B'$$

Let c_i' be the i^{th} row of Σ_{12} . Then

$$E_I(S_{11sij}|s,\underline{x}_2) = \Sigma_{11ij} + c_i'\Sigma_{22}^{-1}(S_{22s} - \Sigma_{22})\Sigma_{22}^{-1} c_j \quad (4.13)$$

From (4.2), (4.7) and (4.10)

$$D_2\rho_i = c_i/\sigma_{1i} \quad (4.14)$$

From (4.13) and (4.14)

$$E_I(S_{11sij}|s,\underline{x}_2) = \Sigma_{11ij} + \sigma_{1i}\sigma_{1j}\rho_i' D_2 \Sigma_{22}^{-1}(S_{22s}-\Sigma_{22})\Sigma_{22}^{-1} D_2 \rho_j$$

$$= \sigma_{1i}\sigma_{1j}(P_{11ij}+\rho_i' P_{22}^{-1} D_2^{-1}(S_{22s}-\Sigma_{22})D_2^{-1} P_{22}^{-1}\rho_j)$$

from (4.1) and (4.12)

$$= \sigma_{1i}\sigma_{1j}(P_{11ij}+\rho_i'\Delta\rho_j) \quad \text{from (4.9)} \qquad (4.15)$$

Hence, since $V_I(S_{11s}|s,\underline{x}_2) = O(n^{-1})$, from (4.3) and (4.15)

$$E_I(R_{11ij}|s,\underline{x}_2) = \sigma_{1i}\sigma_{ij}(P_{11ij}+\rho_i'\Delta\rho_j)/\sigma_{1i}(1+\rho_i'\Delta\rho_i)^{\frac{1}{2}}\sigma_{1j}(1+\rho_j'\Delta\rho_j)^{\frac{1}{2}} + O(n^{-\frac{1}{2}})$$

$$= (P_{11ij}+\rho_i'\Delta\rho_j)/(1+\rho_i'\Delta\rho_i)^{\frac{1}{2}}(1+\rho_j'\Delta\rho_j)^{\frac{1}{2}} + O(n^{-\frac{1}{2}})$$

as required.

The asymptotic expectation of R_{11ij} is the cosine between $(X_1^*)_i$ and $(X_1^*)_j$ in Figure 2.1. Thomson (1951, Ch. 18) notes that:

$$a_i = (1 + \rho_i'\Delta\rho_i)^{\frac{1}{2}}$$

is the ratio of the standard deviations of $(X_1)_i$ in the selected sample and in the original superpopulation. For the case $p_2 = 1$ (when $\Delta = (S_{22s} - \Sigma_{22})/\Sigma_{22})$ and $\Delta < 0$ he sets $b_i = \rho_i(-\Delta)^{\frac{1}{2}}$, which he terms a 'shrinkage factor', and writes

$$E_I(R_{11ij}|s,\underline{x}_2) \doteq \frac{P_{11ij} - b_i b_j}{a_i a_j}$$

He notes that if $S_{22s} = 0$ then $\Delta = -1$ and

$$E_I(R_{11ij}|s,\underline{x}_2) \doteq \frac{P_{11ij} - \rho_i\rho_j}{\sqrt{1-\rho_i^2}\sqrt{1-\rho_j^2}}$$

the usual formula for the partial correlation coefficient between $(X_1)_i$ and $(X_1)_j$ given X_2. This is also, of course, true for the general case of $p_2 > 1$. At the other extreme for the case $p_2 = 1$ we may let $S_{22s} \to \infty$. In this case $\Delta \to \infty$ and

$$E_I(R_{11ij}|s,\underline{x}_2) \to \Delta\rho_i\rho_j/\sqrt{\Delta\rho_i^2}\ \sqrt{\Delta\rho_j^2} = 1$$

i.e. R_{11} approaches the $p_1 \times p_1$ matrix of ones. For the general case $p_2 > 1$ R_{11} will approach a matrix of rank p_2 as we shall see in Section 4.3.

In general it is clear from the example above and from inspection of Figure 2.1, that the effect of selection may be to increase or decrease the P_{11ij}.

Expression (4.6) is still not an easy formula to interpret. Therefore, in the following theorem we assume Δ is small and obtain an approximation. The results of some numerical work (not included) suggest that this approximation is still good for values of Δ as far away from 0 as $-\frac{1}{2}$ or $1(p_2 = 1)$ for a wide range of parameter values.

Theorem 4.2

If Model I holds and $V_I(S_{11s}|s, \underline{x}_2) = O(n^{-1})$ then

$$E_I(R_{11ij}|s,\underline{x}_2) = P_{11ij} + \rho_i'\Delta\rho_j - \tfrac{1}{2}P_{11ij}(\rho_i'\Delta\rho_i+\rho_j'\Delta\rho_j) + O(\Delta^2+n^{-\frac{1}{2}}) \qquad (4.16)$$

Proof

Taking a Binomial expansion of (4.6)

$$E_I(R_{11ij}|s,\underline{x}_2) = (P_{11ij}+\rho_i'\Delta\rho_j)(1-\tfrac{1}{2}\rho_i'\Delta\rho_i)(1-\tfrac{1}{2}\rho_j'\Delta\rho_j) + O(\Delta^2+n^{-\frac{1}{2}})$$

$$= P_{11ij} + \rho_i'\Delta\rho_j - \tfrac{1}{2}P_{11ij}(\rho_i'\Delta\rho_i+\rho_j'\Delta\rho_j) + O(\Delta^2+n^{-\frac{1}{2}})$$

as required.

We may compare (4.16) with the results for covariances in terms of 'relative bias'. From (4.15).

$$E_I\left[(S_{11sij}-\Sigma_{11ij})/\Sigma_{11ij}|s,\underline{x}_2\right] = \rho_i'\Delta\rho_j/P_{11ij} \qquad (4.17)$$

and from (4.16)

$$E_I\left[(R_{11ij}-P_{11ij})/P_{11ij}\,\middle|\,s,\underline{x}_2\right] = \rho_i'\Delta\rho_j/P_{11ij} - \tfrac{1}{2}(\rho_i'\Delta\rho_i+\rho_j'\Delta\rho_j) \qquad (4.18)$$

Both expressions are 'linear' in Δ. For simplicity, consider the case $p_2 = 1$ when $\Delta = (S_{22s}-\Sigma_{22})/\Sigma_{22}$ and the coefficients of Δ in (4.17) and (4.18) are:

$$RB(\Sigma_{11ij}) = \rho_i\rho_j/P_{11ij} \qquad (4.19)$$

$$RB(P_{11ij}) = \rho_i\rho_j/P_{11ij} - \tfrac{1}{2}(\rho_i^2+\rho_j^2) \qquad (4.20)$$

If, as might often be expected to be the case in practice, the sign of $\rho_i\rho_j$ in the same as the sign of P_{11ij}, then the quantities $\rho_i\rho_j/P_{11ij}$ and $-\frac{1}{2}(\rho_i+\rho_j)$ will tend to cancel each other out, especially if P_{11ij} is near unity and ρ_i and ρ_j are not greatly different in magnitude. If P_{11ij} is near zero then both effects will be similar. We give two examples.

Example 4.1: Holt et al (1980b) consider four sets of data with $p_1 = 2$, $p_2 = 1$.

Data set	ρ_1	ρ_2	P_{1112}	$RB(\Sigma_{1112})$	$RB(P_{1112})$
1	.62	.63	.75	.52	-.07
2	.38	.57	.38	.57	.34
3	-.23	.02	.38	-.01	-.04
4	.02	-.16	.22	-.01	-.03

Example 4.2: Gosnell and Schmidt (1936) construct a correlation matrix for voting percentages in a number of areas of Chicago.

	P_{11}					ρ_i
	$(X_1)_2$	$(X_1)_3$	$(X_1)_4$	$(X_1)_5$	$(X_1)_6$	$\underline{X}_2$
$(X_1)_1$	.78	.94	.91	.47	.64	-.62
$(X_1)_2$		.84	.81	.17	.62	-.53
$(X_1)_3$			.96	.40	.62	-.68
$(X_1)_4$				.44	.57	-.64
$(X_1)_5$					.50	-.12
$(X_1)_6$						-.34

(where $(X_1)_1 \ldots (X_1)_6$ are the percentage voting for (1) Smith, (2) Lewis, (3) Roosevelt, (4) Igoe, (5) the Bond Issue, (6) the Wet Vote and X_2 is the median rental value).

We obtain

$RB(\Sigma_{11})$

	$(X_1)_1$	$(X_1)_2$	$(X_1)_3$	$(X_1)_4$	$(X_1)_5$	$(X_1)_6$
$(X_1)_1$	.38	.42	.45	.44	.16	.33
$(X_1)_2$		.28	.43	.42	.37	.29
$(X_1)_3$			.46	.45	.20	.37
$(X_1)_4$				.41	.17	.38
$(X_1)_5$					.01	.08
$(X_1)_6$						.12

$RB(P_{11})$

	$(X_1)_1$	$(X_1)_2$	$(X_1)_3$	$(X_1)_4$	$(X_1)_5$	$(X_1)_6$
$(X_1)_1$	0	.09	.03	.04	-.04	.08
$(X_1)_2$		0	.06	.07	.02	.09
$(X_1)_3$			0	.01	-.04	.08
$(X_1)_4$				0	-.04	.12
$(X_1)_5$					0	.02
$(X_1)_6$						0

Recalling that for $p_2 = 1$ $\Delta = (S_{22}-\Sigma_{22})/\Sigma_{22}$ we may interpret the RB numbers above as follows. If the variance of the design variable is reduced (increased) by A% in selection then Σ_{11ij} is reduced (increased) by $RB(\Sigma_{11ij}) \times A\%$ and P_{11ij} by $RB(P_{11ij}) \times A\%$.

In Example 2, the P_{11ij} are fairly high and the ρ_i are similar and hence, as suggested above, the misspecification effect for the correlation is much smaller than for the covariances. This is also true in Example 1 for data set 1 and to a lesser extent for data set 2. In data sets 3 and 4 one of the ρ_i is near zero. In such cases the effect of selection on the covariances is very small, whereas the effect of selection on the correlations may be greater, since from (4.19) and (4.20)

if $\rho_i = 0$

$$RB(\Sigma_{11ij}) = 0$$

$$RB(P_{11ij}) = -\frac{1}{2}\rho_j^2$$

Even so, in our example both effects are minor.

The Maximum Likelihood Estimator

Note that $\hat{P}_{11}$, defined in (4.4), *is* the MLE of P_1 if (X_1,X_2) are jointly normally distributed by the usual invariance property of MLE's. Note also that $\hat{P}_{11}$ may be expressed, as in Lemma 3.2, as the sample product moment correlation matrix of the variables x^+_{1i} defined in (3.12).

From (3.26), provided $tr(\tilde{S}_{22}\tilde{S}^{-1}_{22s}) = O(1)$ and $V_I(\hat{\Sigma}_{11}|s,x_2) = O(1)$ the conditional asymptotic bias of $\hat{P}_{11}$ as an estimator of P_{11} is zero.

Design-Based Estimator

R^*_{11} is the natural extension of S^*_{11} for estimating P_{11}. The same extension was taken by Koop (1970), Wakimoto (1971c) and Gupta et al. (1978) who, having established S^*_{11s} as a design-unbiased estimator of S_{11}, argued that R^*_{11} was consistent for the finite population analogue of P_{11}.

By comparing Theorems 2.4 and 3.12 we conclude that the expressions for the asymptotic conditional bias of S^*_{11s} will be analogous to the corresponding expressions for S_{11s}. For large n we have by analogy with Theorem 4.1.

$$E_I(R^*_{11ij}|s,\underline{x}_2) \doteq (P_{11ij}+\rho_i'\Delta^*\rho_j)(1+\rho_i'\Delta^*\rho_i)^{-\frac{1}{2}}(1+\rho_j'\Delta^*\rho_j)^{-\frac{1}{2}}$$

where

$$\Delta^* = P_{22}^{-1}(D_2^{-1}S^*_{22s}D_2^{-1}-P_{22})P_{22}^{-1}$$

and for small Δ^* by analogy with Theorem 4.2

$$E_I(R^*_{11ij}|s,\underline{x}_2) \doteq P_{11ij} + \rho_i'\Delta^*\rho_j - \tfrac{1}{2}P_{11ij}(\rho_i'\Delta^*\rho_i+\rho_j'\Delta^*\rho_j)$$

Note that if $p_2 = 1$

$$E_I\left((S^*_{11sij}-\Sigma_{11ij})/\Sigma_{11ij}|s,\underline{x}_2\right) \doteq RB(\Sigma_{11ij})\Delta^*$$

$$E_I\left((R^*_{11ij}-P_{11ij})/P_{11ij}|s,\underline{x}_2\right) \doteq RB(P_{11ij})\Delta^*$$

where $RB(\Sigma_{11ij})$ and $RB(P_{11ij})$ are given in (4.19) and (4.20). These are the same results as for S_{11s} and R_{11} except that Δ^* replaces Δ. This suggests the following intuitive generalisation. The *form* of the misspecification effect for Σ_{11} or P_{11} (as measured by the asymptotic relative conditional bias) is largely determined by the model correlation structure of X_1 and X_2 and is the same for $S_{11s}(R_{11})$ as for $S^*_{11s}(R^*_{11})$, whereas the *degree* of the effect is largely determined by the degree of selection as measured by Δ or Δ^*.

Note that for a given design Δ^* will on average be zero over repeated samples whereas Δ will in general not be.

4.2 Regression Coefficients.

We now partition X_1 into two components which, without loss of generality, we write

$$X_1 = \begin{pmatrix} Y \\ Z \end{pmatrix}$$

where Y is a $p_{11} \times 1$ vector, Z is a $p_{12} \times 1$ vector and $p_{11}+p_{12} = p_1$. Σ_{11}, S_{11s}, $\hat{\Sigma}_{11}$ and S^*_{11s}, defined in (2.4), (2.16), (3.2) and (3.57). respectively are then partitioned conformably as

$$\Sigma_{11} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{pmatrix}, \quad S_{11s} = \begin{pmatrix} S_{yys} & S_{yzs} \\ S_{zys} & S_{zzs} \end{pmatrix}$$

$$\hat{\Sigma}_{11} = \begin{pmatrix} \hat{\Sigma}_{yy} & \hat{\Sigma}_{yz} \\ \hat{\Sigma}_{zy} & \hat{\Sigma}_{zz} \end{pmatrix}, \quad S^*_{11s} = \begin{pmatrix} S^*_{yys} & S^*_{yzs} \\ S^*_{zys} & S^*_{zzs} \end{pmatrix}$$

In this section we consider the estimation of

$$B_{yz} = \Sigma_{yz}\,\Sigma_{zz}^{-1}$$

the marginal regression coefficient matrix of Y on Z. We consider three estimators.

(1) the standared (OLS) estimator:

$$B_{yzs} = S_{yzs}\, S_{zzs}^{-1},$$

(2) the MLE under joint normality of (X_1, X_2):

$$\hat{B}_{yz} = \hat{\Sigma}_{yz}\, \hat{\Sigma}_{zz}^{-1}$$

(3) a design-based estimator

$$B^*_{yzs} = S^*_{yzs}\, S^{*-1}_{zzs}$$

For the case $p_{11} = p_{12} = p_2 = 1$ the MLE was given by Demets and Halperin (1977) and a full discussion of the properties of the three estimators under linear model assumptions was given by Nathan and Holt (1980). Further discussion and empirical study are given by Smith (1982) and Holt et al., (1980b) who also give the MLE in the general multivariate case. The purpose of this section is to extend the basic results of Nathan and Holt (1980) to the general multivariate case. We propose to evaluate the properties of the estimators conditional not only on s and $\underline{x}_2$ but also on $\underline{z}_s = (z_1 \ldots z_n)$. Further unconditional results may then be obtained straightforwardly as in Nathan and Holt (1980).

Throughout this section we shall assume that (X_1, X_2) are jointly normally distributed. We now define some notation. Let

$$\Sigma_{12} = \begin{pmatrix} \Sigma_{yz} \\ \Sigma_{zz} \end{pmatrix}$$

$$B_{y2} = \Sigma_{y2}\Sigma_{22}^{-1},\ B_{z2} = \Sigma_{z2}\ \Sigma_{22}^{-1},\ B_{2z} = \Sigma'_{z2}\Sigma_{zz}^{-1}$$

$$\Sigma_{1.2} = \begin{pmatrix} \Sigma_{yy.2} & \Sigma_{yz.2} \\ \Sigma_{zy.2} & \Sigma_{zz.2} \end{pmatrix}$$

$$\Sigma_{y2.z} = \Sigma_{y2} - \Sigma_{yz}\ \Sigma_{zz}^{-1}\ \Sigma_{z2}$$

$$\Sigma_{22.z} = \Sigma_{22} - \Sigma'_{z2}\ \Sigma_{zz}^{-1}\ \Sigma_{z2}$$

$$B_{y2.z} = \Sigma_{y2.z}\ \Sigma_{22.z}^{-1}$$

$$B_{yz.2} = \Sigma_{yz.2}\ \Sigma_{zz.2}^{-1}$$

$$\Sigma_{y.z2} = \Sigma_{yy} - (\Sigma_{yz}\ \Sigma_{y2}) \begin{pmatrix} \Sigma_{zz} & \Sigma_{z2} \\ \Sigma_{2z} & \Sigma_{zz} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{zy} \\ \Sigma_{2y} \end{pmatrix}$$

$$B_{y(z2)} = (\Sigma_{yz}\ \Sigma_{y2}) \begin{pmatrix} \Sigma_{zz} & \Sigma_{z2} \\ \Sigma_{2z} & \Sigma_{zz} \end{pmatrix}^{-1}$$

We shall use the following identitites

Lemma 4.3

(1) $B_{y(z2)} = (B_{yz.2}\ B_{y2.z})$

(2) $B_{yz} = B_{yz.2} + B_{y2.z}\ B_{2z}$

(3) $B_{y2} = B_{y2.z} + B_{yz.2}\ B_{z2}$

Proof:

We initially consider the geometrical approach of Section 2.1.1. Recall that Y, Z and X_2 define p_{11}, p_{I2} and p_2 vectors respectively in R^{p1+p2}. Let $R(Y)$ be the subspace spanned by the p_{11} vectors of Y etc., and let $R^{\perp}(Y)$ be the orthogonal complement of $R(Y)$. The conditional expectation of Y given (Z, X_2) is

$$B_{y(z2)} \begin{pmatrix} Z \\ X_2 \end{pmatrix}$$

(taking deviations about means) which is represented by the projection of Y onto $R(Z, X_2)$. Now let $Y|X_2$ be the projection of Y onto $R^{\perp}(X_2)$ etc. Then the projection of $Y|X_2$ onto $R(Z|X_2)$ is $B_{yz.2}Z|X_2$ and the projection of $Y|Z$ onto $R(X_2|Z)$ is $B_{y2.z}X_2|Z$. These are depicted in Figure 4.1 where the plane of the paper represents $R(Z,X_2)$.

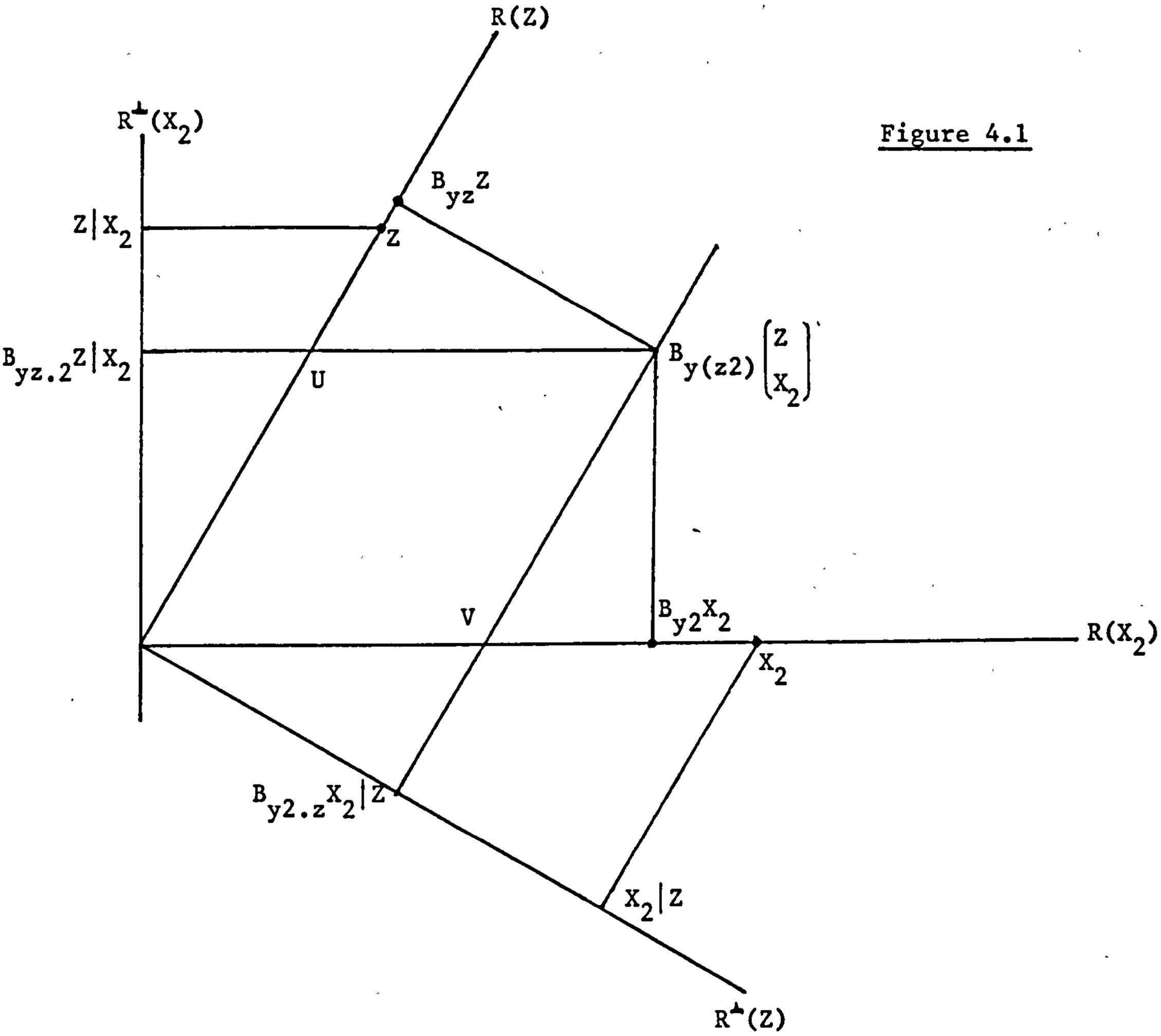


Figure 4.1

Now if $p_{12} = p_2 = 1$ it is clear from a congruent triangles argument that the points U and V in Figure 4.1 are given by $U = B_{yz.2}Z$, $V = B_{y2.z}X_2$ and so

$$B_{y(z2)}\begin{pmatrix} Z \\ X_2 \end{pmatrix} = B_{yz.2}Z + B_{y2.z}X_2$$

and result (1) follows.

For the general case the result also follows because U must be expressible in the form $B_{yz.2}Z|X_2 + AX_2 = B_{yz.2}(Z-BX_2) + AX_2$ which is on R(Z) and hence $U = B_{yz.2}Z$ etc.

To obtain (2) and (3) we note that we may split the projection of Y onto R(Z), say, into two parts: (a) a projection onto $R(Z,X_2)$

and (b) a projection onto $R(Z)$. The projection onto $R(Z.,X_2)$ is given by (1) as $B_{yz.2}Z + B_{y2.z}X_2$ and the projection of this onto $R(Z)$ is $B_{yz.2}Z + B_{y2.z}B_{2z}Z$ which equals $B_{yz}Z$ as required. (3) follows analogously.

(2) and (3) might also, of course, be obtained by conditional expectation arguments. Finally, we given an algebraic derivation.

$$(1)\quad B_{y(z2)} = (\Sigma_{yz}\Sigma_{y2}) \begin{pmatrix}\Sigma_{zz}\Sigma_{z2}\\ \Sigma_{2z}\Sigma_{zz}\end{pmatrix}^{-1}$$

$$= (\Sigma_{yz}\Sigma_{yz}) \begin{pmatrix} \Sigma_{zz.2}^{-1} & -\Sigma_{zz.2}^{-1}\Sigma_{z2}\Sigma_{22}^{-1} \\ -\Sigma_{z2}^{-1}\Sigma_{2z}\Sigma_{zz.2}^{-1}\Sigma_{22}^{-1} & \Sigma_{22}^{-1}\Sigma_{2z}\Sigma_{zz.2}^{-1}\Sigma_{z2}\Sigma_{22}^{-1}\end{pmatrix}$$

(e.g. Morrison, 1976, Ch.2)

$$= (\Sigma_{yz.2}\Sigma_{zz.2}^{-1} \quad \Sigma_{y2}\Sigma_{22}^{-1} - \Sigma_{yz.2}\Sigma_{zz.2}^{-1}\Sigma_{z2}\Sigma_{22}^{-1})$$

$$= \left(B_{yz.2}(\Sigma_{y2} - \Sigma_{yz.2}\Sigma_{zz.2}^{-1}\Sigma_{z2})\Sigma_{22}^{-1}(\Sigma_{22.} - \Sigma_{2z}\Sigma_{zz}^{-1}\Sigma_{z2})\Sigma_{22.z}^{-1}\right)$$

$$= \left(B_{yz.2}\left[\Sigma_{y2} - \left\{\Sigma_{y2}\Sigma_{22}^{-1}\Sigma_{2z} + \Sigma_{yz.2}\Sigma_{zz.2}^{-1}(\Sigma_{zz} - \Sigma_{z2}\Sigma_{22}^{-1}\Sigma_{2z})\right\}\Sigma_{zz}^{-1}\Sigma_{z2}\right]\Sigma_{22.z}^{-1}\right)$$

$$= \left(B_{yz.2}\left[\Sigma_{y2} - \left\{\Sigma_{y2}\Sigma_{22}^{-1}\Sigma_{2z} + \Sigma_{yz} - \Sigma_{y2}\Sigma_{22}^{-1}\Sigma_{2z}\right\}\Sigma_{zz}^{-1}\Sigma_{z2}\right]\Sigma_{22.z}^{-1}\right)$$

$$= \left(B_{yz.2}\left[\Sigma_{y2} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{z2}\right]\Sigma_{22.z}^{-1}\right)$$

$$= (B_{yz.2}\Sigma_{y2.z}\Sigma_{22.z}^{-1})$$

$$= (B_{yz.2}B_{y2.z}) \quad \text{as required}$$

$$(2)\quad B_{yz.2} + B_{y2.z}B_{2z} = \Sigma_{yz.2}\Sigma_{zz.2}^{-1} + \Sigma_{y2.z}\Sigma_{22.z}^{-1}\Sigma_{2z}\Sigma_{zz}^{-1}$$

$$= \left[\Sigma_{yz.2} + \Sigma_{y2.z}\Sigma_{22.z}^{-1}\Sigma_{2z}\Sigma_{zz}^{-1}(\Sigma_{zz} - \Sigma_{z2}\Sigma_{22}^{-1}\Sigma_{2z})\right]\Sigma_{zz.2}^{-1}$$

$$= \left[\Sigma_{yz.2} + \Sigma_{y2.z}\Sigma_{22.z}^{-1}(\Sigma_{22} - \Sigma_{2z}\Sigma_{zz}^{-1}\Sigma_{z2})\Sigma_{22}^{-1}\Sigma_{2z}\right]\Sigma_{zz.2}^{-1}$$

$$= \left[\Sigma_{yz.2} + \Sigma_{y2.z}\Sigma_{22}^{-1}\Sigma_{2z}\right]\Sigma_{zz.2}^{-1}$$

$$= \left[\Sigma_{yz} - \Sigma_{yz}\Sigma_{22}^{-1}\Sigma_{2z} + \Sigma_{yz}\Sigma_{22}^{-1}\Sigma_{2z} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{z2}\Sigma_{22}^{-1}\Sigma_{2z}\right]\Sigma_{zz.2}^{-1}$$

$$= \Sigma_{yz}\Sigma_{zz}^{-1}\left[\Sigma_{zz} - \Sigma_{z2}\Sigma_{22}^{-1}\Sigma_{2z}\right]\Sigma_{zz.2}^{-1}$$

$$= B_{yz} \quad \text{as required}$$

(3) follows similarly.

Now let, as in (2.30),

$$A_y' = (y_1 \ldots y_n)$$

$$A_z' = (z_1 \ldots z_n)$$

Then we may write the three estimators as

$$B_{yzs} = A_y'H_1 \tag{4.21}$$

$$\hat{B}_{yz} = A_y'H_2 \tag{4.22}$$

$$B^*_{yzs} = A_y'H_3 \tag{4.23}$$

where

$$H_1 = P_w A_z S_{zzs}^{-1}/(n-1)$$

and P_w is defined in (2.32)

$$H_2 = MA_z\hat{\Sigma}_{zz}^{-1}/n$$

where $M = M_1 + M_2$, as defined in (3.18) and (3.19)

$$H_3 = M^*A_z S^{*-1}_{zzs}$$

where M* is defined (as M) in Theorem 3.12.

Conditional on $\underline{z}_s$, s and $\underline{x}_2$ these estimators are linear combinations of normal random variables and hence their associated distribution theory is rather easier than that previously concerned with quadratic forms. For this reason we derive both the first and second moments of the estimators, although as before the biases of the estimators will be our main concern in respect of misspecification effects.

The Standard (OLS) Estimator

For the case $n = N = \infty$ we may adopt the geometric approach of Section 2.2.1. Under selection

$$X_2 \to X_2^* = AX_2 \qquad \text{from (2.41)}$$

$$Y \to Y^* = Y - B_{y2}(I-A)X_2$$

$$Z \to Z^* = Z - B_{z2}(I-A)X_2 \qquad \text{from (2.42)}$$

Using Lemma 4.3 (3) we may write

$$Y^* = Y - (B_{y2.z} + B_{yz.2}B_{z2})(I-A)X_2$$

$$= (Y-B_{yz.2}Z - B_{y2.z}X_2) + B_{yz.2}Z^* + B_{y2.z}X_2^* \qquad (4.24)$$

As in the proof of Lemma 4.3, the projection of Y* onto R(Z*) is given by $B_{yzs}Z^*$. This may be taken in two steps. The projections of Y* onto $R(Z^*,X_2^*)$ is from (4.24)

$$B_{yz.2}Z^* + B_{y2.z}X_2^* \qquad (4.25)$$

This is because $Y-B_{yz.2}Z-B_{y2.z}X_2$ is orthogonal to $R(Z,X_2)$ (from Lemma 4.3(1)) and hence is orthogonal to $R(Z^*,X_2^*)$ which is a subspace of $R(Z,X_2)$. Now the projection of (4.25) onto R(Z*) is

$$B_{yz.2}Z^* + B_{y2.z}B_{2zs}Z^*$$

where

$$B_{2zs} = S_{2zs}S_{zzs}^{-1}$$

Hence

$$B_{yzs} = B_{yz.2} + B_{y2.z}B_{2zs}$$

$$= B_{yz} + B_{y2.z}(B_{2zs} - B_{2z}) \quad \text{from Lemma 4.3(2)}$$

This is a generalisation of equation (6.1) of Holt et. al. (1980b). As in the univariate case, $B_{yzs} = B_{yz}$ if Y and X_2 are conditionally independent given Z . The corresponding finite sample results are now given.

Theorem 4.4

If Model I holds and (X_1, X_2) are jointly multivariate normal then

$$E_I(B_{yzs}|\underline{z}_s,s,\underline{x}_2) = B_{yz} + B_{y2.z}(B_{2zs}-B_{2z})$$

$$cov_I(B_{yzsij}, B_{yzsk\ell}|\underline{z}_s, s, \underline{x}_2) = \Sigma_{y.z2ik}S^{-1}_{zzsj\ell}/(n-1)$$

Proof

If (X_1, X_2) are jointly multivariate normal

$$E_I(Y|Z, X_2) = \mu_y + B_{yz.2}(Z-\mu_z) + B_{y2.z}(X_2-\mu_2)$$

where

$$(\mu'_y\mu'_z) = \mu'_1$$

Hence

$$E_I(A'_y|\underline{Z}_s, s, \underline{x}_2) = 1'_n\theta\mu_{y.z2} + B_{yz.2}A'_z + B_{y2.z}A'_2 \qquad (4.26)$$

where

$$\mu_{y.z2} = \mu_y - B_{yz.2}\mu_z - B_{y2.z}\mu_2$$

Hence

$$E_I(B_{yzs}|\underline{z}_s,s,\underline{x}_2) = E_I(A'_y|\underline{z}_s,s,\underline{x}_2)H_1$$

$$= B_{yz.2}A'_zH_1 + B_{y2.z}A'_2H_1$$

since $1'_nP_w = 0$

$$= B_{yz.2} + B_{y2.z}S_{2zs}S^{-1}_{zzs}$$

$$= B_{yz.2} + B_{y2.z}B_{2zs}$$

$$= B_{yz} + B_{y2.z}(B_{2zs}-B_{2z}) \quad \text{from Lemma 4.3 (2)}$$

Now $$B_{yzsij} = ([y_1]_i \cdots [y_n]_i)H_{1j}$$

where $[y_\alpha]_i$ is the i^{th} element of y_α and H_{1j} is the j^{th} column of H_1

$$\mathrm{cov}_I([y_\alpha]_i, [y_\beta]_k | \underline{z}_s, s, \underline{x}_2) = \delta_{\alpha\beta}\Sigma_{y.z2ik}$$

Hence

$$\mathrm{cov}_I(B_{yzsij}, B_{yzsk\ell} | \underline{z}_s, s, \underline{x}_2) = H'_{1j}H_{1\ell}\Sigma_{y.z2ik}$$

$$= [H'_1H_1]_{j\ell}\Sigma_{y.z2ik}$$

$$= \left[S^{-1}_{zzs}A'_zP_wA_zS^{-1}_{zzs}/(n-1)^2\right]_{j\ell}\Sigma_{y.zzjk}$$

$$= S^{-1}_{zzsj\ell}\Sigma_{y.z2ik}/(n-1)$$

The approximate conditional expectation of B_{yzs} gives s and $\underline{x}_2$ may be obtained by taking a Taylor-series expansion

$$E_I(S_{2zs}|s,\underline{x}_2) = S_{22s}B'_{z2}$$

$$E_I(S_{zzs}|s,\underline{x}_2) = \Sigma_{zz} + B_{z2}(S_{22s}-\Sigma_{22})B'_{z2}$$

as in Theorem 2.4.

$$E_I(B_{yzs}|s,\underline{x}_2) \doteq B_{yz} + B_{y2.z}\left[S_{22s}B'_{z2}\left\{\Sigma_{zz} + B_{z2}(S_{22s}-\Sigma_{22})B'_{z2}\right\}^{-1} - B_{2z}\right]$$

a generalisation of equation (2.5) of Nathan and Holt (1980).

The Maximum Likelihood Estimator

Note that $\hat{B}_{yz}$ *is* MLE of B_{yz} under joint normality by the principle of invariance of MLE's. Also $\hat{B}_{yz}$ may be expressed as an OLS estimator in terms of the varibles x^+_{1i} defined in (3.12).

Theorem 4.5

If Model I holds and (X_1, X_2) are jointly multivariate normal then

$$E_I(\hat{B}_{yz} | \underline{z}_s\, s, \underline{x}_2) = B_{yz} + B_{y2.z}(\hat{B}_{2z} - B_{2z})$$

where

$$\hat{B}_{2z} = \tilde{S}_{22} B'_{z2s} \hat{\Sigma}^{-1}_{zz}$$

is the MLE of B_{2z}

and

$$B_{z2s} = S_{z2s} S^{-1}_{22s}$$

$$\mathrm{cov}_I(\hat{B}_{yzij}, \hat{B}_{yzk\ell} | \underline{z}_s, s, \underline{x}_2) = \Sigma_{y.z2ik}\left[\hat{\Sigma}^{-1}_{zz} + \hat{B}'_{2z}(\tilde{S}^{-1}_{22s} - \tilde{S}^{-1}_{22})\hat{B}_{2z}\right]_{j\ell}/n$$

Proof

From (4.26)and (4.22)

$$E_I(\hat{B}_{yz} | \underline{z}_s, s, \underline{x}_2) = \left[1'_n \Theta\, \mu_{y.z2} + B_{yz.2}A'_z + B_{y2.z}A'_2\right]H_2$$

$$= B_{yz.2}A'_z H_2 + B_{y2.z}A'_2 H_2 \qquad (4.27)$$

since

$$1'_n M = 0$$

Now

$$A'_z H_2 = A'_z M A_z \hat{\Sigma}^{-1}_{zz}/n = I_{P_{12}} \qquad (4.28)$$

$$A'_2 H_2 = A'_2 M A_z \hat{\Sigma}^{-1}_{zz}/n$$

and

$$A'_2 M = A'_2 M_1 + A'_2 M_2$$

$$= A'_2 M_2 \quad \text{from (3.24)}$$

$$= n\, \tilde{S}_{22}(A'_2 P_w A_2)^{-1} A'_2 P_w$$

$$\therefore \quad A_2'H_2 = n\,\tilde{S}_{22}(A_2'P_wA_2)^{-1}A_2'P_wA_z\hat{\Sigma}_{zz}^{-1}/n$$

$$= \tilde{S}_{22}S_{22s}^{-1}S_{2zs}\hat{\Sigma}_{zz}^{-1} \qquad (4.29)$$

Now

$$B_{2z} = \Sigma_{2z}\Sigma_{zz}^{-1}$$

$$= \Sigma_{22}(\Sigma_{z2}\Sigma_{22}^{-1})'\Sigma_{zz}^{-1}$$

$$= \Sigma_{22}B_{z2}'\Sigma_{zz}^{-1}$$

Hence from Theorem 3.1 (or see Smith, 1982) the MLE of B_{2z} is

$$\hat{B}_{2z} = \tilde{S}_{22}S_{22s}^{-1}\tilde{S}_{2zs}\hat{\Sigma}_{zz}^{-1}$$

as given in the Theorem. Hence from (4.29)

$$A_2'H_2 = \hat{B}_{2z}$$

and substituting into (4.27) using (4.28)

$$E_I(\hat{B}_{yz}|\underline{z}_s,s,\underline{x}_2) = B_{yz.2} + B_{y2.z}\hat{B}_{2z}$$

$$= B_{yz} + B_{y2.z}(\hat{B}_{2z}-B_{2z}) \text{ from Lemma 4.3(2)}$$

As in Theorem 4.4

$$\mathrm{cov}_I(\hat{B}_{yzij},\hat{B}_{yzk\ell}|\underline{z}_s,s,\underline{x}_2) = [H_2'H_2]_{j\ell}\Sigma_{y.z2ik}$$

and

$$H_2'H_2 = \hat{\Sigma}_{zz}^{-1}\,A_z'M^2A_z\hat{\Sigma}_{zz}^{-1}/n^2$$

Now

$$A_z'M^2A_z = A_z'MA_z + A_z'(M_2^2-M_2)A_z \quad \text{from (3.32)}$$

$$= n\left[\hat{\Sigma}_{zz}+B_{z2s}(\tilde{S}_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}-\tilde{S}_{22})B_{z2s}'\right]$$

$$H_2'H_2 = \left[\hat{\Sigma}_{zz}^{-1} + \hat{\Sigma}_{zz}^{-1}B_{z2s}(\tilde{S}_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}-\tilde{S}_{22})B_{z2s}'\right]/n$$

$$= \left[\hat{\Sigma}_{zz}^{-1} + \hat{B}_{2z}'(\tilde{S}_{22s}^{-1}-\tilde{S}_{22}^{-1})\hat{B}_{2z}\right]/n$$

and the result follows.

It follows that $\hat{B}_{yz}$ is asymptotically unbiased. For B_{z2s} is asymptotically unbiased for B_{z2} under selection as in conventional regression and hence $\hat{B}_{2z}$ is asymptotically unbiased for $\Sigma_{22}B'_{z2}\Sigma_{zz}^{-1} = B_{2z}$

The Design-based Estimator

Theorem 4.6

If Model I holds and (X_1, X_2) are jointly multivariate normal then

$$E_I(B^*_{yzs}|\underline{z}_s, s, \underline{x}_2) = B_{yz} + B_{y2.z}(B^*_{2zs} - B_{2z})$$

$$\mathrm{cov}_I(B^*_{yzsij}, B^*_{yzsk\ell}|\underline{z}_s, s, \underline{x}_2) = \Sigma_{y.z2ik}\ \mathrm{tr}(M^{*2})\left[S^{*-1}_{zzs}S^{**}_{zzs}S^{*-1}_{zzs}\right]_{j\ell}$$

where S^{**}_{zzs} is defined as in Theorem 3.12

Proof

From (4.23) and (4.26)

$$E_I(B^*_{yzs}|\underline{z}_s, s, \underline{x}_2) = \left[1'_n \otimes \mu_{y.z2} + B_{yz.2}A'_z + B_{y2.z}A'_2\right]H_3$$

$$= B_{yz.2}A'_zH_3 + B_{y2.z}A'_2H_3$$

since

$$1'_nM^* = 0$$

Now

$$A'_zH_3 = A'_zM^*A_zS^{*-1}_{zzs} = I_{p_{12}}$$

$$A'_2H_3 = A'_2M^*A_zS^{*-1}_{zzs}$$

$$= S^*_{2zs}S^{*-1}_{zzs} = B^*_{2zs}$$

where

$$S^*_{2zs} = \left[\sum_{i<j}(x_{2i}-x_{2j})(z_i-z_j)'/\pi_{ij}\right]/\left[\sum_{\alpha\neq\beta}1/\pi_{\alpha\beta}\right]$$

Hence

$$E_I(B^*_{yzs}|z_s,s,\underline{x}_2) = B_{yz.2} + B_{y2.z}B^*_{2zs}$$

$$= B_{yz} + B_{y2.z}(B^*_{2zs}-B_{2z}) \quad \text{from Lemma 4.3(2)}$$

The second order result follows as in Theorems 4.4 and 4.5 by noting that

$$H_3'H_3 = S^{*-1}_{zzs}A_z'M^*M^*A_zS^{*-1}_{zzs}$$

$$= tr(M^{*2})S^{*-1}_{zzs}S^{**}_{zzs}S^{*-1}_{zzs}$$

from the proof of Theorem 3.12.

4.3 Principal Components Analysis

In this section we consider a principal components analysis of the aggregate covariance matrix, Σ_{11} (and to a limited extent the correlation matrix, P_{11}, defined in (4.1)). Note that a disaggregated within-group principal components analysis might proceed as in Krzanowski (1979), and that Tortora (1980) suggests an alternative within - strata approach using dummy variables.

Let $\gamma_1 \ldots \gamma_p$ be normalised eigenvectors of Σ_{11} corresponding to eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$, i.e.

$$\Sigma_{11}\gamma_i = \lambda_i\gamma_i \qquad i = 1 \ldots p \tag{4.30}$$

$$\gamma_i'\gamma_j = \delta_{ij}, \quad \text{the Kronecker } \delta \tag{4.31}$$

The i^{th} principal component of Σ_{11} is $y_i = \gamma_i'X_1$. In principal components analysis we might be interested in various aspects of the γ_i and λ_i (e.g. Mardia et al., 1979, Ch.8). Here we only consider the point estimation of the γ_i and λ_i. As in Sections 4.1 and 4.2, we consider three estimators:

(1) Standard estimators - the eigenvalues and eigenvectors of S_{11s} (or R_{11} defined in (4.3),

(2) MLE's - the eigenvalues and eigenvectors of $\hat{\Sigma}_{11}$ (or $\hat{P}_{11}$ defined in (4.4)),

(3) design-based estimators - the eigenvalues and eigenvectors of S^*_{11s} (or R^*_{11} defined in (4.5)).

We shall need to approximate eigenvalues and eigenvectors as functions of Σ_{11} etc. The following result is due to Girschick (1939) (who used the alternative normalisation $\gamma_i'\gamma_i = \lambda_i$). Higher order terms in the expansion are given by Lawley (1956), Mallows (1961), Watern aux (1976) and Sugiura (1976). (See also Sibson, 1979, for a non-stochastic approach). The'proof' follows the intuitive approach of Girschick.

Lemma 4.7:

Let Σ be a $p \times p$ non-negative definite matrix with simple eigenvalues $\lambda_1 > \ldots > \lambda_p$ and corresponding eigenvectors $\gamma_1 \ldots \gamma_p$. Let $S = \Sigma + d\Sigma$ be a random covariance matrix with eigenvalues $\ell_1 > \ldots > \ell_p$ (a.s) and corresponding eigenvectors $g_1 \ldots g_p$ such that $d\Sigma = O_p(n^{-\frac{1}{2}})$. Then

$$\ell_i = \lambda_i + \gamma_i' d\Sigma \gamma_i + O_p(n^{-1}) \tag{4.32}$$

$$g_i = \gamma_i + \sum_{j \neq i} w_{ij}\gamma_j + O_p(n^{-1}) \tag{4.33}$$

where

$$w_{ij} = \gamma_j' d\Sigma \gamma_i / (\lambda_i - \lambda_j)$$

'Proof'

$$\Sigma\gamma_i = \lambda_i\gamma_i \tag{4.34}$$

$$\gamma_i'\gamma_j = \delta_{ij} \tag{4.35}$$

Let

$$d\lambda_i = \ell_i - \lambda_i$$

$$d\gamma_i = g_i - \gamma_i$$

Then differentiating (4.34) gives

$$d\Sigma\gamma_i + \Sigma d\gamma_i = d\lambda_i\gamma_i + \lambda_i d\gamma_i \qquad (4.36)$$

Multiply (4.36) by γ_i'

$$\gamma_i' d\Sigma\gamma_i + \gamma_i'\Sigma d\gamma_i = d\lambda_i\gamma_i'\gamma_i + \lambda_i\gamma_i' d\gamma_i$$

From (4.34) and (4.35)

$$\gamma_i' d\Sigma\gamma_i + \lambda_i\gamma_i' d\gamma_i = d\lambda_i + \lambda_i\gamma_i' d\gamma_i$$

and (4.32) follows

Now multiply (4.36) by γ_j' where $j\neq i$ and use (4.34) and (4.35)

$$\gamma_j' d\Sigma\gamma_i + \lambda_j\gamma_j' d\gamma_i = \lambda_i\gamma_j' d\gamma_i$$

$$\therefore \qquad \gamma_j' d\gamma_i = \gamma_j' d\Sigma\gamma_i/(\lambda_i-\lambda_j) \qquad j\neq i \qquad (4.37)$$

Differentiating (4.35) gives

$$\gamma_i' d\gamma_i = 0 \qquad (4.38)$$

The spectral decomposition of Σ is

$$\Sigma = \sum_j \lambda_j\gamma_j\gamma_j' \qquad (4.39)$$

Combining (4.37) - (4.39)

$$\Sigma d\gamma_i = \sum_{j\neq i} \lambda_j\gamma_j\gamma_j' d\Sigma\gamma_i/(\lambda_i-\lambda_j)$$

But

$$\Sigma^{-1} = \Sigma\lambda_j^{-1}\gamma_j\gamma_j'$$

Hence

$$d\gamma_i = \sum_k \sum_{j\neq i} \lambda_k^{-1}\gamma_k\gamma_k'\lambda_j\gamma_j\gamma_j' d\Sigma\gamma_i/(\lambda_i-\lambda_j)$$

$$= \sum_{j\neq i} \gamma_j\gamma_j' d\Sigma\gamma_i/(\lambda_i-\lambda_j) \qquad \text{from (4.35)}$$

as required

Corollary 4.8

To a first-order approximation

$$E(\ell_i) = \lambda_i + \gamma_i'\big(E(S) - \Sigma\big)\gamma_i$$

$$\operatorname{cov}(\ell_i,\ell_j) = \sum_{\alpha\beta\bar{\alpha}\bar{\beta}} \gamma_{i\alpha}\gamma_{i\beta}\gamma_{j\bar{\alpha}}\gamma_{j\bar{\beta}} \operatorname{cov}(S_{\alpha\beta},S_{\bar{\alpha}\bar{\beta}})$$

$$E(g_i) = \gamma_i + \sum_{j\neq i} \gamma_j'\Big[E(S) - \Sigma\Big]\gamma_i\gamma_j/(\lambda_i-\lambda_j)$$

$$\operatorname{cov}(g_i,g_j) = \sum_{\alpha\neq i} \sum_{\beta\neq j} \operatorname{cov}(w_{i\alpha},w_{j\beta})\gamma_\alpha\gamma_\beta'$$

where

$$\operatorname{cov}(w_{i\alpha},w_{j\beta}) = \sum_{k\ell mn}\gamma_{ik}\gamma_{\alpha\ell}\gamma_{jm}\gamma_{\beta n} \operatorname{cov}(S_{k\ell},S_{mn})/(\lambda_i-\lambda_\alpha)(\lambda_j-\lambda_\beta)$$

Proof:

This follows directly from Lemma 4.7.

Corollary 4.9

If $E(S) = \Sigma$, $\operatorname{cov}(S_{ij},S_{k\ell}) = (\Sigma_{ik}\Sigma_{j\ell}+\Sigma_{i\ell}\Sigma_{jk})/n$ (as in the standard IID normal case) then to a first-order approximation

$$E(\ell_i) = \lambda_i \qquad V(\ell_i) = 2\lambda_i^2/n \qquad \operatorname{cov}(\ell_i,\ell_j) = 0 \quad i \neq j$$

$$E(g_i) = 0 \qquad V(g_i) = \lambda_i \sum_{\alpha\neq i}\lambda_\alpha\gamma_\alpha\gamma_\alpha'/(\lambda_i-\lambda_\alpha)^2 n$$

$$\operatorname{cov}(g_i,g_j) = -\lambda_i\lambda_j\gamma_j\gamma_i'/(\lambda_i-\lambda_j)^2 n \qquad i \neq j$$

Proof:

From Corollary 4.8

$$E(\ell_i) = \lambda_i + \gamma_i'(\Sigma-\Sigma)\gamma_i = \lambda_i \qquad \text{as required}$$

$$V(\ell_i) = \sum_{\alpha\beta\bar{\alpha}\bar{\beta}} \gamma_{i\alpha}\gamma_{i\beta}\gamma_{j\bar{\alpha}}\gamma_{j\bar{\beta}}(\Sigma_{\alpha\bar{\alpha}}\Sigma_{\beta\bar{\beta}} + \Sigma_{\alpha\bar{\beta}}\Sigma_{\beta\bar{\alpha}})/n$$

$$= 2(\gamma_i'\Sigma\gamma_i)^2/n = 2\lambda_i^2/n \quad \text{as required}$$

Similarly

$$\text{cov}(\ell_i,\ell_j) = 2(\gamma_i'\Sigma\gamma_j)^2/n = 0$$

$$E(g_i) = \gamma_i \quad \text{as required}$$

$$\text{cov}(w_{i\alpha},w_{i\beta}) = \sum_{k\ell mn} \gamma_{ik}\gamma_{\alpha\ell}\gamma_{im}\gamma_{\beta n}(\Sigma_{km}\Sigma_{\ell n} + \Sigma_{kn}\Sigma_{\ell m})/(\lambda_i-\lambda_\alpha)(\lambda_i-\lambda_\beta)n$$

$$= (\gamma_i'\Sigma\gamma_i\gamma_\alpha'\Sigma\gamma_\beta + \gamma_i'\Sigma\gamma_\beta\gamma_\alpha'\Sigma\gamma_i)/(\lambda_i-\lambda_\alpha)(\lambda_i-\lambda_\beta)n$$

$$= (\lambda_i\lambda_\alpha\delta_{\alpha\beta} + \lambda_i^2\delta_{\alpha i}\delta_{\beta i})/(\lambda_i-\lambda_\alpha)(\lambda_i-\lambda_\beta)n$$

$$\therefore \quad V(g_i) = \sum_{\alpha\neq i} \lambda_i\lambda_\alpha\gamma_\alpha\gamma_\alpha'/(\lambda_i-\lambda_\alpha)^2 n$$

Similarly

$$\text{cov}(w_{i\alpha},w_{j\beta}) = (\lambda_i\lambda_\alpha\delta_{ij}\delta_{\alpha\beta} + \lambda_i\lambda_j\delta_{i\beta}\delta_{j\alpha})/(\lambda_i-\lambda_\alpha)(\lambda_j-\lambda_\beta)n$$

If $i\neq j$ $\text{cov}(g_i,g_j) = \lambda_i\lambda_j\gamma_j\gamma_i'/(\lambda_i-\lambda_j)(\lambda_j-\lambda_i)n$

$$= -\lambda_i\lambda_j\gamma_j\gamma_i'/(\lambda_i-\lambda_j)^2 n$$

Corollary 4.9 gives the standard first-order asymptotic results for IID normal samples (e.g. Girschick, 1939; Anderson, 1963). Corresponding results for IID non-normal samples (e.g. Davis, 1977) would also follow from Corollary 4.8. Finite sample theory for the standard IID case is very intractable e.g. James, (1960, 1964), Johnson and Kotz (1972, ch. 39) and no exact expressions for moments appear to be available except for some symmetric functions of the eigenvalues e.g. $\text{tr}(\Sigma) = \Sigma\lambda_i$ (Krishnaiah and Chattopadhyay, 1975; Mathai , 1980). Even the extension of the first-order approximation theory to the case of multiple eigenvalues is complicated (e.g. Anderson, 1963) and no simple expressions for moments are available. The quality of the first-order approximations in IID sampling is not obviously very good from the limited finite sample investigations that have been

carried out (e.g. Bebbington and Smith, 1977; Waternaux, 1976). Anderson (1963) notes, for example, that ℓ_1 will always be upwardly biased for λ_1 in finite samples. Also, for example, the asymptotic normality of the ℓ_i (which we do not use) seems unlikely to be a good approximation in small samples since the ℓ_i perform like sample variances. Some further work on asymptotic theory is in Muirhead (1978), Fujikoshi (1980), Krishnaiah and Lee (1979) and Tyler (1981).

The Standard Estimators

The first rather obvious comment is that if rank $(\Sigma_{11}) = m < p_1$ then rank$(S_{11s}) \leq m$ (in fact rank $(S_{11s}) = m$ a.s) i.e. selection cannot increase the number of principal components (unlike factor analysis; see Section 4.4). Let us consider the case $n = N = \infty$ and the geometrical approach of Section 2.2.1. The principal components of Σ_{11} can be taken as a natural orthogonal basis for the subspace of $R^{p_1+p_2}$ spanned by X_1. For let $\Sigma_{11} = \Gamma\Lambda\Gamma'$ be the spectral decomposition of Σ_{11} and let $M = \Lambda^{\frac{1}{2}}\Gamma'$. Then the columns of M augmented by p_2 zeros may be taken as the co-ordinates of $(X_1)_1, \ldots (X_1)_{p_1}$ in $R^{p_1+p_2}$ for the inner-product of these p_1 vectors is $M'M = \Gamma\Lambda\Gamma' = \Sigma_{11}$ as required. The i^{th} principal component of Σ_{11} is $y_i = X_1'\gamma_i$ which is represented by $\begin{pmatrix} M \\ O \end{pmatrix}\gamma_i$ which is a vector with $\sqrt{\lambda_i}$ in the i^{th} position and zeros elsewhere, i.e. the principal components lie along the axes of the co-ordinate system.

In general, the effect of selection will be to map the y_i onto non-orthogonal vectors and hence to change the principal component structure.

Lemma 4.10

If $n = N = \infty$ then the principal components of S_{11s} are the same as the principal components of Σ_{11} (with possibly different variances) iff

$$b_i'(S_{22s} - \Sigma_{22})b_j = 0 \qquad i \neq j$$

where $b_i = \gamma_i'B$ is the (row) vector of coefficients of the linear regression of y_i on X_2.

e.g. if (a) $b_i = 0 \quad i \neq i_o$ for some i_o

or (b) $S_{22s} = \Sigma_{22}$

Proof: Necessity:

If S_{11s} has the same principal components as Σ_{11} then $\Gamma'S_{11s}\Gamma$ is diagonal where $\Gamma = (\gamma_1 \ldots \gamma_{p_1})$. But from (2.38)

$$S_{11s} = \Sigma_{11} + B(S_{22s}-\Sigma_{22})B'$$

Hence $\Gamma'B(S_{22s}-\Sigma_{22})B'\Gamma$ must be diagonal. The ij^{th} element of this matrix is $b_i'(S_{22s}-\Sigma_{22})b_j$ and so the result follows:

Sufficiency:

If $b_i'(S_{22s}-\Sigma_{22})b_j = 0 \quad i \neq j$ then $\Gamma'S_{11s}\Gamma$ is diagonal, say

$$\Gamma'S_{11s}\Gamma = \text{diag}(\ell_i)$$

$$\therefore \quad S_{11s}\Gamma = \Gamma \text{diag}(\ell_i)$$

$$\therefore \quad S_{11s}\gamma_i = \ell_i\gamma_i$$

and the $\gamma_1 \ldots \gamma_{p1}$ are the eigenvectors of S_{11s} and hence S_{11s} and Σ_{11} have the same principal components.

Case (b) above, where only one of the b_i is non-zero, is of most interest. This might occur, for example, if X_1 was a vector of expenditure variables and X_2 was an income variable which was related to the first principal component but not to the others. Note that the eigenvalues of a diagonal matrix are equal to the elements on the diagonal and so if (b) holds the eigenvalues of S_{11s} are

$$\ell_i = \lambda_{i_o} + b_{i_o}(S_{22s}-\Sigma_{22})b_{i_o}' \qquad i = i_o$$

$$= \lambda_i \qquad i \neq i_o$$

Hence the only effect of selection is to increase or reduce a single eigenvalue. This might of course be misleading to the Model II researcher since the eigenvalues are usually taken to measure the relative importance of the different components.

In general we shall need to use Lemma 4.7 to approximate the eigenvalues and eigenvectors of S_{11s}. However, one other example where an exact result (for $n = N = \infty$) is available is:

Example 4.3 : $\Sigma_{11} = \sigma^2 I$, $p_2 = 1$

Here

$$\lambda_i = \sigma^2 \qquad i = 1....p_1$$

$\gamma_1 \ldots \gamma_{p_1}$ are any orthonormal basis of R^{p_1}.

From (2.38) we may write:

$$S_{11s} = \sigma^2(I - BB'/B'B) + \left\{(S_{22s}-\Sigma_{22})B'B + \sigma^2\right\}BB'/B'B$$

But $I - BB'/B'B$ and $BB'/B'B$ are orthogonal projection matrices and so the eigenvalues of S_{11s} are

$$\ell_1 = \sigma^2 + (S_{22s}-\Sigma_{22})B'B \qquad \text{(assuming } S_{22s} \geq \Sigma_{22}\text{)}$$

$$\ell_i = \sigma^2 \qquad i \neq 1$$

and the eigenvectors are

$$g_1 = B/\sqrt{B'B}$$

$g_2 \ldots g_{p_1}$ are any orthonormal basis of the (p_1-1)-dimensional subspace of R^{p_1} spanned by the columns of $I - BB'/B'B$.

This is an example of the discontinuity problems that can occur in the 'eigenvector function' when the eigenvalues of Σ_{11} are equal or nearly equal, in which case a small perturbation in Σ_{11} can drastically affect the eigenvectors.

Before considering approximation results we consider the most extreme possible effects of selection (when $n = N = \infty$). Recall from (2.38)

$$S_{11s} = \Sigma_{11} + B(S_{22s}-\Sigma_{22})B'$$

At one extreme $S_{22s} = 0$ and $S_{11s} = \Sigma_{1.2} = \Sigma_{11} - B\Sigma_{22}B'$, the partial covariance matrix of X_1 given X_2. Hence if we know S_{22s} is 'smaller' than Σ_{22} then we might in practice compute the principal component structure of the partial covariance matrix of X_1 given X_2 and, if it is not very different from that of S_{11s}, not worry too much about the effect of selection (assuming 'continuity'). Note that the same remarks hold for the principal components of the correlation matrix.

At the other extreme as $S_{22s} \to \infty$, $S_{11s} \to BS_{22}B'$ which is a matrix of rank p_2. This is likely to have a very different principal component structure to that of Σ_{11}. Note that R_{11} (defined in 4.3) $\to \left[\rho_i'\Delta\rho_j(\rho_i'\Delta\rho_i)^{-\frac{1}{2}}(\rho_j'\Delta\rho_j)^{-\frac{1}{2}}\right]$ as $S_{22s} \to \infty$. If, say $p_2 = 1$, $R_{11} \to \left[\rho_i\rho_j/\rho_i\rho_j\right] = \left[1\right]$ i.e. the matrix of ones.

Theorem 4.11

Suppose that the eigenvalues $\lambda_1 > \ldots > \lambda_{p_1}$ of Σ_{11} are simple. Let $\ell_1 \geq \ldots \geq \ell_{p_1}$ be the eigenvalues of S_{11s} and $g_1 \ldots g_{p_1}$ be the corresponding eigenvectors. Then providing $S_{11s} - \Sigma_{11}$ is 'small'

$$E_I(\ell_i | s, \underline{x}_2) \doteq \lambda_i + b_i(S_{22s} - \Sigma_{22})b_i'$$

$$E_{II}(\ell_i | s, \underline{x}_2) \doteq \lambda_i$$

$$E_I(g_i | s, \underline{x}_2) \doteq \gamma_i + \sum_{j \neq i} w_{ji}\gamma_j$$

where

$$w_{ji} = b_j(S_{22s} - \Sigma_{22})b_i'/(\lambda_i - \lambda_j)$$

$$E_{II}(g_i | s, \underline{x}_2) \doteq \gamma_i$$

Proof

From Theorem 2.4:

$$E_I(S_{11s} | s, \underline{x}_2) \doteq \Sigma_{11} + B(S_{22s} - \Sigma_{22})B'$$

Hence from Corollary 4.8:

$$E_I(\ell_i|s,\underline{x}_2) \doteq \lambda_i + \gamma_i' B(S_{22s} - \Sigma_{22})B'\gamma_i$$

$$= \lambda_i + b_i(S_{22s}-\Sigma_{22})b_i'$$

and

$$E_I(g_i|s,\underline{x}_2) \doteq \gamma_i + \sum_{j\neq i} w_{ji}\gamma_j$$

where

$$w_{ji} = \gamma_j' B(S_{22s}-\Sigma_{22})B\gamma_i/(\lambda_i-\lambda_j)$$

$$= b_j(S_{22s}-\Sigma_{22})b_i'/(\lambda_i-\lambda_j)$$

The results for Model II follow as special cases. The result for ℓ_i is a natural extension of Theorem 2.4 viewing ℓ_i as a sample variance of a linear combination of the X_1 variables. The result for g_i is less easy to interpret but note that the absence of γ_i in the 'misspecification effect' is due to the normalisation $\gamma_i'\gamma_i = 1$ which means that γ_i lies on the surface of a hypersphere and forces $d\gamma_i$ to be orthogonal to $\gamma_i(\gamma_i' d\gamma_i = 0)$. Note also that the term $(\lambda_i-\lambda_j)^{-1}$ means that g_i is very unstable if λ_i is close to any of the other eigenvalues (c.f. the standard IID case).

Simulation

~~Similar~~ results for the case $p_2 = 1$ suggest that the approximations in Theorem 4.11 are very good for large samples, especially for ℓ_i, for wide ranges of values of S_{22s} (say $0.5\Sigma_{22}<S_{22s}<2\Sigma_{22}$). Note that bounds on ℓ_i and g_i in the case $n = N = \infty$ may be obtained from Wilkinson (1965 p.104) without using limiting arguments.

For the case $p_2 = 1$ let $r_i = \text{corr}_I(y_i, X_2)$

Then

$$r_i = \frac{\gamma_i' \Sigma_{12}}{\sqrt{\lambda_i \Sigma_{22}}}$$

$$b_i = \gamma_i'\Sigma_{12}/\Sigma_{22} = r_i\sqrt{\lambda_i/\Sigma_{22}}$$

Let $\Delta = (S_{22s}-\Sigma_{22})/\Sigma_{22}$ as in (4.9)

Then

$$E_I(\ell_i|s,\underline{x}_2) \doteq \lambda_i + b_i^2\Delta\Sigma_{22}$$

$$= \lambda_i + r_i^2\lambda_i\Delta$$

$$= \lambda_i(1+r_i^2\Delta) \qquad (4.40)$$

Hence the selection effect on ℓ_i depends on r_i and on Δ. Note that

$$E_I(\ell_i/\Sigma\ell_i \mid s, \underline{x}_2) \doteq \lambda_i(1+r_i^2\Delta)/\Sigma\lambda_i(1+r_i^2\Delta)$$

Hence the proportion of variance explained by the i^{th} principal component is approximately unaffected by selection if $r_1 = \ldots = r_{p_1}$

We now consider the principal components of the correlation matrix. Let $\tilde{X}_1$ be the standardised vector X_1 i.e.

$$\tilde{X}_1 = D_1^{-1}X_1$$

where $$D_1 = \text{diag}(\sigma_{1i})$$

σ_{1i} is defined in (4.2)

Let $\lambda_1^P \geq \ldots \geq \lambda_{p_1}^P$ be the eigenvalues of P_{11} (defined in 4.1) with corresponding eigenvectors $\gamma_1^P \ldots \gamma_{p_1}^P$. Let $\tilde{B}$ be the regression coefficient matrix of $\tilde{X}_1$ on X_2

i.e. $$\tilde{B} = D_1^{-1}B$$

and let b_i^P be the (row) vector of coefficients of the regression of $\gamma_i^{P\prime}\tilde{X}_1$ on X_2

i.e. $$b_i^P = \gamma_i^{P\prime}\tilde{B}$$

Theorem 4.12

Suppose the eigenvalues of P_{11} are simple i.e. $\lambda_1^P > \ldots > \lambda_{p_1}^P$. Let $\ell_1^P \geq \ldots \geq \ell_{p_1}^P$ be the eigenvalues of R_{11} (defined in 4.3) and $g_1^P \ldots g_{p_1}^P$ be the corresponding eigenvectors. Then providing $S_{11s} - \Sigma_{11}$ is 'small'

$$E_I(\ell_i^P \mid s, \underline{x}_2) \doteq \lambda_i^P + b_i^P(S_{22s}-\Sigma_{22})b_i^{P\prime} - \lambda_i^P \sum_\alpha (\gamma_i^P)_\alpha^2 \left(\tilde{B}(S_{22s}-\Sigma_{22})\tilde{B}'\right)_{\alpha\alpha}$$

$$E_{II}(\ell_i^P \mid s, \underline{x}_2) \doteq \lambda_i^P$$

$$E_I(g_i^P \mid s, \underline{x}_2) \doteq \gamma_i^P + \sum_{j\neq i} w_{ji}\gamma_j^P$$

where

$$w_{ji} = \left[b_j^P(S_{22s}-\Sigma_{22})b_i^{P\prime} - \frac{1}{2}(\lambda_i^P+\lambda_j^P)\sum_\alpha(\gamma_i^P)_\alpha(\gamma_j^P)_\alpha\left[\tilde{B}(S_{22s}-\Sigma_{22})\tilde{B}'\right]_{\alpha\alpha}\right]/(\lambda_i^P-\lambda_j^P)$$

Proof:

From Theorem 4.2, if $S_{22s}-\Sigma_{22}$ is small

$$E_I(R_{11ij}|s,\underline{x}_2) \doteq P_{11ij} + \rho_i'\Delta\rho_j - \frac{1}{2}P_{11ij}(\rho_i'\Delta\rho_i + \rho_j'\Delta\rho_j)$$

Hence from Corollary 4.8

$$E_I(\ell_i^P|s,\underline{x}_2) = \lambda_i^P + \sum_{\alpha\beta}(\gamma_i^P)_\alpha(\gamma_i^P)_\beta\left[\rho_\alpha'\Delta\rho_\beta - \frac{1}{2}P_{11\alpha\beta}(\rho_\alpha'\Delta\rho_\alpha + \rho_\beta'\Delta\rho_\beta)\right] \qquad (4.41)$$

Now

$$b_i^P = \gamma_i^{P\prime}\tilde{B}$$

$$= \gamma_i^{P\prime}D_1^{-1}B$$

$$= \gamma_i^{P\prime}D_1^{-1}\Sigma_{12}\Sigma_{22}^{-1}$$

$$= \gamma_i^{P\prime}D_1^{-1}\Sigma_{12}D_2^{-1}P_{22}^{-1}D_2^{-1}$$

where P_{22} is defined in (4.12)

$$= \gamma_i^{P\prime}\begin{pmatrix}\rho_1'\\ \vdots\\ \rho_{P_1}'\end{pmatrix}P_{22}^{-1}D_2^{-1}$$

$$= \sum_\alpha(\gamma_1^P)_\alpha\rho_\alpha'P_{22}^{-1}D_2^{-1}$$

Hence

$$b_i^P(S_{22s}-\Sigma_{22})b_i^{P\prime} = \sum_\alpha(\gamma_i^P)_\alpha\rho_\alpha'P_{22}^{-1}(D_2^{-1}S_{22s}D_2^{-1}-P_{22})P_{22}^{-1}\sum_\beta(\gamma_i^P)_\beta\rho_\beta$$

$$= \sum_{\alpha\beta}(\gamma_i^P)_\alpha(\gamma_i^P)_\beta\rho_\alpha'\Delta\rho_\beta \quad \text{from (4.9)}$$

Similarly

$$\left(\tilde{B}(S_{22s}-\Sigma_{22})\tilde{B}'\right)_{\alpha\beta} = \rho_\alpha'\,\Delta\rho_\beta$$

Hence from (4.41)

$$E_I(\ell_i^p|s,\underline{x}_2) = \lambda_i^p + b_i^p(S_{22s}-\Sigma_{22})b_i^{p\prime} - \frac{1}{2}\Big[\sum_\alpha(\gamma_i^p)_\alpha(P_{11}\gamma_i^p)_\alpha\rho_\alpha'\Delta\rho_\alpha + \sum_\beta(\gamma_i^p)_\beta(P_{11}\gamma_i^p)_\beta\rho_\beta'\Delta\rho_\beta\Big]$$

$$= \lambda_i^p + b_i^p(S_{22s}-\Sigma_{22})b_i^{p\prime} - \lambda_i^p\sum_\alpha(\gamma_i^p)_\alpha^2\,\rho_\alpha'\Delta\rho_\alpha$$

since

$$P_{11}\gamma_j^p = \lambda_i^p\gamma_i^p$$

$$= \lambda_i^p + b_i^p(S_{22s}-\Sigma_{22})b_i^{p\prime} - \lambda_i^p\sum_\alpha(\gamma_i^p)_\alpha^2\,\left(\tilde{B}(S_{22s}-\Sigma_{22})\tilde{B}'\right)_{\alpha\alpha}$$

as required.

Similarly from Corollary 4.8

$$E_I(g_i^p|s,\underline{x}_2) \doteq \gamma_i^p + \sum_{j\neq i} w_{ji}\gamma_j^p$$

where

$$w_{ji} = \gamma_j^{p\prime}(R_{11}-P_{11})\gamma_i^p/(\lambda_i^p-\lambda_j^p)$$

$$= \sum_{\alpha\beta}(\gamma_j^p)_\alpha(\gamma_i^p)_\beta\Big[\rho_\alpha'\Delta\rho_\beta - \frac{1}{2}P_{11\,\alpha\beta}(\rho_\alpha'\Delta\rho_\alpha + \rho_\beta'\Delta\rho_\beta)\Big]/(\lambda_i^p-\lambda_j^p)$$

$$= b_j^p(S_{22s}-\Sigma_{22})b_i^{p\prime} - \frac{1}{2}\Big[\sum_\alpha(\gamma_j^p)_\alpha\lambda_i^p(\gamma_i^p)_\alpha\rho_\alpha'\Delta\rho_\alpha + \sum_\beta(\gamma_i^p)_\beta\lambda_j^p(\gamma_j^p)_\beta\rho_\beta'\Delta\rho_\beta\Big]/(\lambda_i^p-\lambda_j^p)$$

$$= b_j^p(S_{22s}-\Sigma_{22})b_i^{p\prime} - \frac{1}{2}(\lambda_i^p+\lambda_j^p)\sum_\alpha(\gamma_i^p)_\alpha(\gamma_j^p)_\alpha \left[\tilde{B}(S_{22s}-\Sigma_{22})\tilde{B}'\right]_{\alpha\alpha}/(\lambda_i^p-\lambda_j^p)$$

The results for Model II follow as special cases.

Note that the first two terms in the expectations of ℓ_i^p and g_i^p correspond to the expectations of ℓ_i and g_i in Theorem 4.11. In order to compare the two results consider the case $p_2 = 1$:

Let

$$r_i^P = \text{corr}_I(\gamma_i^P, \tilde{X}_1, X_2)$$

$$= b_i^P \sqrt{\Sigma_{22}/\lambda_i^P}$$

$$\tilde{r}_i = \text{corr}_I((\tilde{X}_1)_i, X_2)$$

$$= \tilde{B}_i \sqrt{\Sigma_{22}}$$

$$\Delta = (S_{22s} - \Sigma_{22})/\Sigma_{22}$$

Then the relative biases of ℓ_i and ℓ_i^P (c.f. Section 4.1) are

$$E_I\left[(\ell_i - \lambda_i)/\lambda_i \,|\, s, \underline{x}_2\right] \doteq r_i^2 \Delta \qquad \text{from (4.40)}$$

$$E_I\left[(\ell_i^P - \lambda_i^P)/\lambda_i^P \,|\, s, \underline{x}_2\right] \doteq (r_i^{P2} - \sum_\alpha (\gamma_i^P)_\alpha^2 \tilde{r}_\alpha^2)\Delta$$

Numerical values from Example 4.2 are

i	1	2	3	4	5	6
r_i^2	.416	.035	.020	.000	.027	.011
r_i^{P2}	.368	.061	.019	.037	.008	.017
$r_i^{P2} - \sum_\alpha (\gamma_i^P)_\alpha^2 \tilde{r}_\alpha^2$	.040	-.021	-.154	-.226	-.374	-.422

Note that the values r_i^2 and r_i^{P2} are similar and that the relative bias for the correlation matrix is smaller for the first two components but takes a relatively large value for the last four components. This effect, however, is less worrying if we consider the absolute biases.

i	1	2	3	4	5	6
$\lambda_i r_i^2$	252.7	1.9	0.5	0.0	0.2	0.1
$\lambda_i^P (r_i^{P2} - \sum_\alpha (\gamma_i^P)_\alpha^2 \tilde{r}_\alpha^2)$	0.17	-0.02	-0.07	-0.03	-0.03	-0.01

Hence in this example the proportions of variance explained by the different components will be much more affected by selection for the covariance matrix than for the correlation matrix.

We might summarise the selection effect on the eigenvectors by the Euclidean distance between $g_i(g_i^p)$ and $\gamma_i(\gamma_i^p)$ (cf. Bebbington and Smith, 1977). From Theorem 4.11 for the case $p_2 = 1$:

$$E_I\left[\{(g_i-\gamma_i)'(g_i-\gamma_i)\}^{\frac{1}{2}} \mid s, \underline{x}_2\right] \doteq d_i\Delta$$

where $d_i = |b_i|\{\sum_{j\neq i} b_j^2/(\lambda_i-\lambda_j)^2\}^{\frac{1}{2}}\Sigma_{22}$

(assuming the variance of g_i is negligible) and from Theorem 4.12:

$$E_I\left[\{(g_i^p-\gamma_i^p)'(g_i^p-\gamma_i^p)\}^{\frac{1}{2}} \mid s, \underline{x}_2\right] \doteq e_i^p\Delta$$

where

$$e_i^p = \{\sum_{j\neq i}\left[b_j^p b_i^p - \frac{1}{2}(\lambda_i^p+\lambda_j^p)\sum_\alpha(\gamma_i^p)_\alpha(\gamma_j^p)_\alpha\,\tilde{B}_\alpha^2\right]^2/(\lambda_i^p-\lambda_j^p)^2\}^{\frac{1}{2}}$$

We now evaluate d_i and e_i^p for Example 4.2 and for comparison we also evaluate

$$d_i^p = |b_i^p|\{\sum_{j\neq i} b_j^{p2}/(\lambda_i^p-\lambda_j^p)^2\}^{\frac{1}{2}}$$

i	1	2	3	4	5	6
d_i	.047	.007	.043	.008	.044	.034
d_i^p	.096	.105	.066	.051	.036	.027
e_i^p	.021	.038	.060	.107	.087	.025

As for the eigenvalues the second term in e_i^p tends to cancel out the first term making e_i^p usually less than d_i^p. In this example all the selection effects are fairly small but this is because the first principal component is very dominating and selection occurs mainly along the first component. In other examples, the selection effects may be much larger.

The above numerical work has been aimed at investigating the conjecture that selection effects are smaller for correlation matrices than for covariance matrices. Although the data from Example 4.2 provides some evidence in support of this conjecture we would suggest that in general it would be dangerous to take this as an assumption. Note that we have demonstrated that it can be quite misleading to assume that R_{11} is a covariance matrix from a distribution where the $(X_.)_i$ have unit variances and then to apply the results of Theorem 4.11.

In order to obtain the variances of the estimates, let the conditional covariance matrix of the principal components $y_1 \ldots y_{p_1}$ given X_2 be

$$\Sigma_{y.2} = \left[\delta_{ij}\lambda_i - b_i\Sigma_{22}b_j'\right]$$

As in Theorem 2.10 and Corollary 2.11 let

$$\phi_{ys} = \left[b_i S_{22s} b_j'\right]$$

$$\Sigma_y^* = \Sigma_{y.2} + \phi_{ys}$$

Theorem 4.13

Under the conditions of Theorem 4.11 and the assumption that $X_1|X_2$ is multivariate normal:

$$V_I(\ell_i|s,\underline{x}_2) \doteq 2(\Sigma^{*2}_{yii} - \phi^2_{ysii})/n$$

$$V_{II}(\ell_i|s,\underline{x}_2) \doteq 2\lambda_i^2/n$$

$$\mathrm{cov}_I(\ell_i,\ell_j|s,\underline{x}_2) \doteq 2(\Sigma^*_{yij} - \phi^2_{ysij})/n$$

$$\mathrm{cov}_{II}(\ell_i,\ell_j|s,\underline{x}_2) \doteq 0$$

Proof:

Combining Corollaries 2.11 and 4.8:

$$\text{cov}_I(\ell_i,\ell_j|\ s,\underline{x}_2) = \sum_{\alpha\beta}\sum_{\bar{\alpha}\bar{\beta}}\gamma_{i\alpha}\gamma_{i\beta}\gamma_{j\bar{\alpha}}\gamma_{j\bar{\beta}}(\Sigma^*_{\alpha\bar{\alpha}}\Sigma^*_{\beta\bar{\beta}} + \Sigma^*_{\alpha\bar{\beta}}\Sigma^*_{\beta\bar{\alpha}} - \phi_{s\alpha\bar{\alpha}}\phi_{s\beta\bar{\beta}} - \phi_{s\alpha\bar{\beta}}\phi_{s\beta\bar{\alpha}})/n$$

$$= 2(\gamma_i'\Sigma^*\gamma_j)^2/n - 2(\gamma_i'\phi_s\gamma_j)/n$$

$$= 2(\Sigma^{*2}_{yij} - \phi^2_{ysij})/n$$

as required

The results for Model II follow as a special case with $\phi_{ys} = 0$, $\Sigma^*_y = \text{diag}(\lambda_i)$.

Note that, as remarked after Theorem 2.8, even under balanced sampling with $S_{22s} = \Sigma_{22}$ the result for Model I differs from that of Model II (which is the same as that given in Corollary 4.9). This is because of the non-centrality introduced by conditioning on $\underline{x}_2$. If, for example, we take an srs design and evaluate the moments of ℓ_i under Model I then they would be the same as for Model II in Theorem 4.13.

Note also that the variances and covariances of the ℓ_i are analogous to those of the diagonal elements of S_{11s} in Corollary 2.11. As in Theorem 2.10 we may rewrite the main result in Theorem 4.13 as:

$$\text{cov}_I(\ell_i,\ell_j) = (2\Sigma^2_{y.2ij} + 4\Sigma_{y.2ij}\phi_{ysij})/n$$

Note finally that $\text{cov}_I(\ell_i,\ell_j)$ is linear in S_{22s} since the quadratic terms cancel each other out.

Theorem 4.14

Under the conditions of Theorem 4.11 and the assumption that $X_1|X_2$ is multivariate normal:

$$\text{cov}_I(g_i,g_j|s,\underline{x}_2) = \sum_{\alpha\neq i}\sum_{\beta\neq j}\Big[\ \Sigma^*_{yij}\Sigma^*_{y\alpha\beta} + \Sigma^*_{yi\beta}\Sigma^*_{yj\alpha} - \phi_{ysij}\phi_{ys\alpha\beta} - \phi_{ysi\beta}\phi_{ysj\alpha}\Big]$$

$$\gamma_\alpha\gamma_\beta'/n(\lambda_i-\lambda_\alpha)(\lambda_j-\lambda_\beta)$$

$$\text{cov}_{II}(g_i, g_j | s, \underline{x}_2) = \lambda_{i\alpha \neq i}^{\Sigma} \lambda_\alpha \gamma_\alpha \gamma_\alpha' / (\lambda_i - \lambda_\alpha)^2 n \quad \text{if} \quad i = j$$

$$= -\lambda_i \lambda_j \gamma_j \gamma_i' / (\lambda_i - \lambda_j)^2 n \quad \text{if} \quad i \neq j$$

Proof:

From Corollary 4.8:

$$\text{cov}_I(g_i, g_j | s, \underline{x}_2) = \sum_{\alpha \neq i}^{\Sigma} \sum_{\beta \neq j}^{\Sigma} \text{cov}_I(w_{i\alpha}, w_{j\beta} | s, \underline{x}_2) \gamma_\alpha \gamma_\beta'$$

where

$$\text{cov}_I(w_{i\alpha}, w_{j\beta} | s, \underline{x}_2) = \sum_{k\ell mn} \gamma_{ik} \gamma_{\alpha\ell} \gamma_{jm} \gamma_{\beta n} \text{cov}(S_{11sk\ell}, S_{11smn} | s, \underline{x}_2) / (\lambda_i - \lambda_\alpha)(\lambda_j - \lambda_\beta)$$

$$= \sum_{k\ell mn} \gamma_{ik} \gamma_{\alpha\ell} \gamma_{jm} \gamma_{\beta n} (\Sigma^*_{km} \Sigma^*_{\ell n} + \Sigma^*_{kn} \Sigma^*_{\ell m} - \phi_{skm} \phi_{s\ell n} - \phi_{skn} \phi_{s\ell m}) / n(\lambda_i - \lambda_\alpha)(\lambda_j - \lambda_\beta)$$

$$= (\gamma_i' \Sigma^* \gamma_j \gamma_\alpha' \Sigma^* \gamma_\beta + \gamma_i' \Sigma^* \gamma_\beta \gamma_\alpha' \Sigma^* \gamma_j - \gamma_i' \phi_s \gamma_j \gamma_\alpha' \phi_s \gamma_\beta - \gamma_i' \phi_s \gamma_\beta \gamma_\alpha' \phi_s \gamma_j) / n(\lambda_i - \lambda_\alpha)(\lambda_j - \lambda_\beta)$$

$$= (\Sigma^*_{yij} \Sigma^*_{y\alpha\beta} + \Sigma^*_{yi\beta} \Sigma^*_{yj\alpha} - \phi_{ysij} \phi_{ys\alpha\beta} - \phi_{ysi\beta} \phi_{ysj\alpha}) / n(\lambda_i - \lambda_\alpha)(\lambda_j - \lambda_\beta) \quad \text{as required}$$

The result for Model II follows as a special case with $\phi_{ys} = 0$, $\Sigma^*_y = \text{diag}(\lambda_i)$.

The Maximum Likelihood Estimator

If the eigenvalues of Σ_{11} are simple then the eigenvalues $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_{p_1}$ of $\hat{\Sigma}_{11}$ and the corresponding eigenvectors $\hat{\gamma}_1 \ldots \hat{\gamma}_{p_1}$ *are* the MLE's of $\lambda_1 \ldots \lambda_p$ and $\gamma_1 \ldots \gamma_p$ respectively by the principle of invariance for ML estimation.

Theorem 4.15

If the eigenvalues of Σ_{11} are simple and $\hat{\Sigma}_{11} - \Sigma_{11} = O_p(n^{-1})$

(e.g. if $X_1|X_2$ is multivariate normal) then

$$E_I(\hat{\lambda}_i|s,\underline{x}_2) = \lambda_i + b_i(\tilde{S}_{22}-\Sigma_{22})b_i' + \left[tr(\tilde{S}_{22}\tilde{S}_{22s}^{-1})-p_2-1\right]\Sigma_{y.2ii}/n+O_p(n^{-1})$$

$$E_I(\hat{\gamma}_i|s,\underline{x}_2) = \gamma_i + \sum_{j\neq i} w_{ji}\gamma_j + O_p(n^{-1})$$

where

$$w_{ji} = \left[b_j(\tilde{S}_{22}-\Sigma_{22})b_i' + \left(tr(\tilde{S}_{22}\tilde{S}_{22s}^{-1}) - p_2 - 1\right)\Sigma_{y.2ij}/n\right]/(\lambda_i-\lambda_j)$$

Proof:

This follows by combining (3.26) and Corollary 4.8.

Theorem 4.16:

If the eigenvalues of Σ_{11} are simple and $X_1|X_2$ is multivariate normal then

$$cov_I(\hat{\lambda}_i,\hat{\lambda}_j|s,\underline{x}_2) = 2n'\Sigma_{y.2ij}^2/n^2 + 4\Sigma_{y.2ij}\psi_{yij}/n + O_p(n^{-3/2})$$

$$cov_I(\hat{\gamma}_i,\hat{\gamma}_j|s,\underline{x}_2) = \sum_{\alpha\neq i}\sum_{\beta\neq j}\left\{n'\left[\Sigma_{y.2ij}\Sigma_{y.2\alpha\beta} + \Sigma_{y.2i\beta}\Sigma_{y.2j\alpha}\right]/n\right.$$

$$\left. + \Sigma_{y.2ij}\psi_{y\alpha\beta} + \Sigma_{y.2\alpha\beta}\psi_{yij} + \Sigma_{y.2i\beta}\psi_{yj\alpha} + \Sigma_{y.2j\alpha}\psi_{yi\beta}\right\}$$

$$/n(\lambda_i-\lambda_\alpha)(\lambda_j-\lambda_\beta) + O_p(n^{-3/2})$$

where

$$n' = n - p_2 - 1 + tr(\tilde{S}_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}\tilde{S}_{22s}^{-1})$$

$$\psi_{yij} = b_i\tilde{S}_{22}\tilde{S}_{22s}^{-1}\tilde{S}_{22}b_j'$$

Proof:

This follows from Theorem 3.5 and Corollary 4.8 as in the proofs of Theorems 4.13 and 4.14.

Note that, assuming the p-convergence of S_{22s} away from zero,

$$E_I(\hat{\lambda}_i|s,\underline{x}_2) = \lambda_i + O_p(n^{-1})$$

$$E_I(\hat{\gamma}_i|s,\underline{x}_2) = \gamma_i + O_p(n^{-1})$$

Hence the second moments of $\hat{\lambda}_i$ and $\hat{\gamma}_i$ dominate their MSE's asymptotically. Note that S_{22s} features as a denominator in the covariances of $\hat{\lambda}_i$ and $\hat{\gamma}_i$ and so these variances will increase as S_{22s} 'decreases', since B is then estimated more poorly.

The Design-Based Estimator

Theorem 4.17

If the eigenvalues of Σ_{11} are simple and $\ell_1^* \geq \ldots \geq \ell_{p_1}^*$ and $g_1^* \ldots g_{p_1}^*$ are eigenvalues and eigenvectors of S_{11s}^*, then providing $S_{11s}^* - \Sigma_{11}$ is 'small'

$$E_I(\ell_i^*|s,\underline{x}_2) \doteq \lambda_i + b_i(S_{22s}^* - \Sigma_{22})b_i'$$

$$E_I(g_i^*|s,\underline{x}_2) \doteq \gamma_i + \sum_{j\neq i} w_{ji}\gamma_j$$

where $$w_{ji} = b_j(S_{22s}^* - \Sigma_{22})b_i'/(\lambda_i - \lambda_j)$$

Proof: This follows from Theorem 3.12 and Corollary 4.8.

This result is analogous to Theorem 4.11. Simulation results suggest that the approximation is good for large samples for a wide range of values of S_{22s}^*. Note that Theorem 4.12 would also apply in the design based case if we substitute the eigenvalues and eigenvectors of R_{11}^* (defined in (4.5)) for ℓ_i^p and g_i^b and substitute S_{22s}^* for S_{22s}. The discussion following Theorems 4.11 and 4.12 would also apply if we substitute Δ^* for Δ. As noted in Section 4.1, the form of the misspecification-effect is the same for S_{11s} or S_{11s}^* and is largely determined by the model correlation structure whereas the degree of the misspecification effects depends largely on the differences $S_{11s} - \Sigma_{11}$ and $S_{11s}^* - \Sigma_{11}$.

Theorem 4.18

If the eigenvalues of Σ_{11} are simple and $X_1|X_2$ is multivariate normal then providing $S^*_{11s} - \Sigma_{11}$ is 'small'

$$\mathrm{cov}_I(\ell^*_i,\ell^*_j|s,\underline{x}_2) \doteq 2\mathrm{tr}(M^2)\left[\Sigma^2_{y.2ij} + 2\Sigma_{y.2ij}\psi^*_{yij}\right]$$

$$\mathrm{cov}_I(g^*_i,g^*_j|s,\underline{x}_2) \doteq \mathrm{tr}(M^2)\sum_{\alpha\neq i}\sum_{\beta\neq j}\left[\Sigma_{y.2ij}\Sigma_{y.2\alpha\beta}+\Sigma_{y.2i\beta}\Sigma_{y.2j\alpha}+\Sigma_{y.2ij}\psi^*_{y\alpha\beta}\right.$$

$$\left.+\Sigma_{y.2\alpha\beta}\psi^*_{yij}+\Sigma_{y.2i\beta}\psi^*_{yj\alpha}+\Sigma_{y.2j\alpha}\psi^*_{yi\beta}\right]/(\lambda_i-\lambda_\alpha)\ (\lambda_j-\lambda_\beta)$$

where M is defined in Theorem 3.12

and

$$\psi^*_{yij} = b_i S^{**}_{22s} b'_j$$

Proof:

This follows from Theorem 3.12 and Corollary 4.8.

4.4 Factor Analysis

The study of the effect of sample selection on factor analysis has a long history. This is probably because of an interest in the conjecture that there exists a fundamental invariant structure of human abilities which ought to be reflected in the factor analysis of mental test data for any group of human subjects.

Thomson (1938) and Ledermann (1938b), adopting Pearson's (1903) population-level approach, supposed that (X_1,X_2) jointly obeyed a factor analysis model in the population and that selection was made according to X_2. They showed that under 'univariate selection', i.e. $p_2 = 1$, the new *correlation matrix* also obeyed a factor analysis model with the same number of common factors but with different factor loadings and communalities. Thomson and Ledermann (1939) then showed that under 'multivariate selection', i.e. $p_2 > 1$, the new correlation matrix also obeyed a factor analysis model but with, in general, p_2 extra common factors, where X_1 did not depend on the new 'selection factors'. It is interesting to note that Thomson and

Ledermann made use of the concept of a superpopulation of 'might have beens' (Thomson and Ledermann, 1939, p.289) of which the sample and the population might be thought of as samples. Their reason for adopting this approach was, however, somewhat different from ours. They were concerned with the process of evolution where one generation (the sample) was selected from a previous generation (the population) and where the sample might even outnumber the population.

Thomson and Ledermann's results were considered as rather disturbing since they cast doubt on the idea that factors can be 'interpreted as basic and identifiable psychological processes' (Thurstone, 1945; see also Lawley and Maxwell, 1971, p.114). A more optimistic view was expressed by Thurstone (1945) who showed that, although the factor loadings were generally altered, zero loadings were not ('simple structure' is invariant) and he argued that any new 'selection factors' should be 'classed with the residual factors which reflect the conditions of particular experiments' (p.179). Thomson (1951, p.304) found this argument rather unconvincing.

A rather elegant review of the work of Thomson, Ledermann and Thurstone is given by Ahmavaara (1954) who used the geometric approach of Section 2.2.1. He showed also that the effect of multivariate selection on factor analysis based on the *covariance matrix* was just to change the columns of the factor loading matrix of X_1 proportionately.

Meredith (1964a) took an approach which is closer to our interests. He supposed that in the population X_1 obeyed a factor analysis model and that selection took place according to other variables X_2 which were independent of the unique factors of the X_1 model and were related to the common factors as in Lawley (1943) (i.e. as in our Model I). He showed that selection did not introduce any new common factors and that, for a given rotation, the new factor loadings were unchanged for the covariance matrix (Ahmavaara's (1954) result differed because he imposed a particular normalisation on the factors). Our discussion will be based on Meredith's work.

Finally, Bloxom (1972) extended Meredith's work by allowing the selection variables X_2 also to be correlated with the unique factors. This complicates the picture greatly and in particular introduces new common factors as in Thomson and Ledermann (1939).

In this section we shall make Meredith's assumption that X_2 is independent of the unique factors. This seems to us reasonable since the common factors are intended to tap all the 'behavioural' components of the X_1 variables and the unique factors represent measurement error and the 'non-behavioural' unique components of the X_1 variables (e.g. Fielding, 1977). This is not to say that Bloxom's approach could not be adapted to our problem but we suspect that a more interesting extension would be to assume that the *variances* of the unique factors depend on X_2 e. g. the reliability of attitude items might vary between social classes or age groups.

Suppose then that X_1 obeys a factor analysis model (marginally) in the super population, i.e.

$$X_1 = \mu_1 + \Lambda f + u$$

where f is an unobserved $m \times 1$ vector of common factors

u is an unobserved $p_1 \times 1$ vector of unique factors

$$E(f) = 0 \quad , \quad E(u) = 0$$

$$V(f) = \Phi, \ \mathrm{cov}(f, u) = 0 \quad , \quad V(u) = \Psi = \mathrm{diag}\ (\psi_i)$$

The covariance structure of X_1 is

$$\Sigma_{11} = \Lambda\Phi\Lambda' + \Psi \tag{4.42}$$

Let $D_1 = \mathrm{diag}(\sigma_{1i})$ where σ_{1i} is defined in (4.2)
Then

$$P_{11} = D_1^{-1}\Sigma_{11}D_1^{-1} = \Lambda_p \ \Phi \ \Lambda_p' + \Psi_p \tag{4.43}$$

where

$$\Lambda_p = D_1^{-1}\Lambda, \quad \Psi_p = D_1^{-1} \ \Psi \ D_1^{-1}$$

Λ is the matrix of (covariance) factor loadings.

Λ_p is the matrix of (correlation) factor loadings.

Suppose that X_2 is independent of u and that f is related to X_2 by Model I of section 2.1., i.e.

$$E(f|X_2) = B_{f2}(X_2-\mu_2)$$

$$V(f|X_2) = \Sigma_{f.2}$$

Then for the case $n = N = \infty$ the selected covariance matrix is from (2.38):

$$S_{11s} = \Sigma_{11} + B(S_{22s}-\Sigma_{22})B'$$

$$= \Lambda\Phi_s\Lambda' + \Psi \qquad (4.44)$$

where

$$\Phi_s = \Phi + B_{f2}(S_{22s}-\Sigma_{22})B'_{f2} \qquad (4.45)$$

is the selected covariance matrix of f ,

since

$$B = \Lambda B_{f2} \qquad (4.46)$$

Let

$$D_{1s} = \text{diag}(S^{\frac{1}{2}}_{11sii})$$

Then the selected correlation matrix is

$$R_{11} = D^{-1}_{1s}S_{11s}D^{-1}_{1s} = \Lambda_R\Phi_s\Lambda^{-1}_R + \Psi_R \qquad (4.47)$$

where

$$\Lambda_R = D^{-1}_{1s}\Lambda \quad , \quad \Psi_R = D^{-1}_{1s}\ \Psi D^{-1}_{1s}$$

Comparing (4.42) and (4.44) we see that S_{11s} obeys a factor analysis model with the same number of common factors and the same unique variances as Σ_{11} and, given the rotation in (4.44), the same factor loadings. On the other hand, comparing (4.43) and (4.47), we see that R_{11} obeys a factor analysis model with the same number of common factors as P_{11} but with different unique variances and factor loadings (unless $D_{1s} \propto I$). Hence the recommendation that under selection it is best to work with the X_1 variables in their original units. Note, however, that each zero in Λ_p has a zero in the corresponding position in Λ_R , i.e. 'simple structure' is preserved.

We have shown that, given a particular rotation of the model for Σ_{11}, there exists a rotation of the model for S_{11s} such that the factor loadings are the same. It is straightforward to show (Meredith, 1964a) that given *any* rotation of the model for Σ_{11} there exists a rotation of the model for S_{11} such that the factor loadings are the same. Conversely, it is also true that given any rotation of the model for S_{11s} there exists a rotation of the model for Σ_{11} such that the factor loadings are invariant. The problem for interpretation, however, is that in general $\Phi_s \neq \Phi$. This means, for example, (1) that although we might obtain the 'correct' factor loading matrix from S_{11s} we might be misled into thinking that the common factors are correlated when in fact (in the population) they are not or (2) that we might be misled by the 'percentages of variance' explained by each factor and by the communalities. Meredith cannot deal with this problem since he supposes that the data refers to several samples selected from the population but where no population data is available, Meredith (1964b) therefore only considers the problem of appropriately rotating within-sample factor analyses in order to obtain an invariant Λ with generally varying Φ_s matrices.

We suppose, however, that (finite) population data is available on X_2. For the remainder of this section we consider six different methods of estimating Λ, Φ and Ψ. As usual there is the rotation-indeterminacy of the factor analysis model. We shall therefore set $\Phi = I$ and accept that we can only estimate Λ up to an orthogonal rotation.

1. The Standard Estimator

Suppose that we enter S_{11s} into a standard factor analysis package, say using ML estimation. From (4.44) we shall obtain consistent estimators of Ψ and of a rotation of Λ. In general we obtain an inconsistent estimator of $\Lambda\Phi\Lambda'$. Note that if $m = 1$ the estimated factor loading vector will be consistent for a vector proportional to Λ.

2. A Design-based Estimator

Suppose that we enter S^*_{11s} into a standard factor analysis package. Then we shall obtain ξ-consistent estimators of Ψ and of a rotation of Λ. The estimator of $\Lambda\Phi\Lambda'$ will be $p\xi$-consistent.

The estimators are in general inefficient.

3. Exact ML Estimator

Assume that (X_1, X_2) are jointly normally distributed. Recall from Section (3.2) that the likelihood may be expressed as

$$p(\underline{x}_{1s}|\underline{x}_2, s, \mu_{1.2}, \Sigma_{1.2}, B)p(s|\underline{x}_2)\ p(\underline{x}_2|\mu_2, \Sigma_{22}) \qquad (4.48)$$

Unlike principal components analysis we cannot maximise (4.48) by entering $\hat{\Sigma}_{11}$ into a ML factor analysis package (see estimator 4) because Λ, Φ and Ψ are not 1-1 functions of Σ_{11}. Instead the parameters in (4.48) are restricted by

$$B = \Lambda B_{f2}\ ,\ \Sigma_{1.2} = \Lambda\Phi\Lambda' + \Psi - \Lambda B_{f2}\Sigma_{22}^{-1}B'_{f2}\Lambda'$$

For the purposes of ML estimation it is inconvenient to set $\Phi = I$ and so we adopt an alternative identifying restriction which may be removed later by rotation. We may naturally parameterise the model by $(\mu_{1.2}, \Lambda, B_{f2}, \Sigma_{f.2}, \Psi, \mu_2, \Sigma_{22})$

where

$$\Sigma_{f.2} = \Phi - B_{f2}\Sigma_{22}B'_{f2}$$

since

$$\Sigma_{1.2} = \Lambda\Sigma_{f.2}\Lambda' + \Psi$$

One obvious source of underidentification occurs under the rotation $\Lambda^* = \Lambda H\ ,\ B^*_{f2} = H^{-1}B_{f2},\ \Sigma^*_{f.2} = H^{-1}\Sigma_{f.2}H'^{-1}$ where H is a non-singular $m \times m$ matrix. For in this case the likelihood is unchanged by substituting $(\mu_{1.2}, \Lambda^*, B^*_{f2}, \Sigma^*_{f.2}, \Psi, \mu_2, \Sigma_{22})$. We propose therefore to impose the constraint:

$$\Sigma_{f.2} = I_m$$

If $m = 1$ this removes the underidentification (up to sign indeterminacy in Λ and B_{f2}). If $m > 1$ then there is still a lack of identification because H may be an orthogonal matrix. This may be removed by requiring that $\Lambda'\Psi^{-1}\Lambda$ is a diagonal matrix with elements on the diagonal arranged in descending order of magnitude (c.f. Lawley and Maxwell, 1971, p.8). Let us now rewrite the likelihood, ignoring the middle term in (4.48) which is fixed, as

$$p(\underline{x}_{1s}|\underline{x}_{2s}, \mu_{1.2}, \Lambda, B_{f2}, \Psi)\ p(\underline{x}_2|\mu_2, \Sigma_{22}) \tag{4.49}$$

It seems reasonable to assume that $(\mu_{1.2}, \Lambda, B_{f2}, \Psi)$ and (μ_2, Σ_{22}) are Cartesian independent (Definition 1.1). Hence the likelihood may be maximised by maximising the two terms of (4.49) separately. The second term is maximised as before by

$$\hat{\mu}_2 = \bar{x}_2\ ,\ \hat{\Sigma}_{22} = \tilde{S}_{22} \tag{4.50}$$

The first term is, for the case $m = 1$, the likelihood of the multiple-indicator multiple cause (MIMIC) model of Jöreskog and Goldberger (1975). In their model the X_2 variables were causes of the latent variable f and they were specifically interested in B_{f2} (whereas we treat X_2 as a 'nuisance'). They show how to maximise the first term of our likelihood and note the generalisability of their approach to the case $m > 1$. They contrast the case where $\underline{x}_{2s}$ is fixed with the case where $x_{21} \ldots x_{2n}$ are a random sample from $N_{p_2}(\mu_2, \Sigma_{22})$. In the latter case the likelihood is

$$p(\underline{x}_{1s}|\underline{x}_{2s}, \mu_{1.2}, \Lambda, B_{f2}, \Psi)\ p(\underline{x}_{2s}|\mu_2, \Sigma_{22}) \tag{4.51}$$

They note that the MLE's of $(\mu_{1.2}, \Lambda, B_{f2}, \Psi)$ are the same in both cases. But (4.51) is the likelihood of a random sample of n values (x_{1i}, x_{2i}) from a multivariate normal distribution with

$$E\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \mu_{1.2} + \Lambda B_{f2}\mu_2 \\ \mu_2 \end{pmatrix}\ ,\ V\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \Lambda(B_{f2}\Sigma_{22}B'_{f2}+I)\Lambda'+\psi & \Lambda B_{f2}\Sigma_{22} \\ \Sigma_{22}B'_{f2}\Lambda' & \Sigma_{22} \end{pmatrix} \tag{4.52}$$

which may be maximised for example using the programme LISREL (Joreskog and Sörbom, 1978) to give MLE's $(\hat{\mu}_{1.2}, \hat{\Lambda}, \hat{B}_{f2}, \hat{\Psi})$. Note that (4.51) is maximised when $\mu_2 = \bar{x}_{2s}$ and $\Sigma_{22} = S_{22s}$ but these values should be disregarded in favour of the actual MLE's in (4.44). Note also that the mean structure of X_1 is effectively unconstrained in (4.52) so

$$\therefore \qquad \mu_{1.2} + \Lambda B_{f2} \bar{x}_{2s} = \bar{x}_{1s}$$

$$\hat{\mu}_1 = \bar{x}_{1s} + \hat{\Lambda} \hat{B}_{f2}(\bar{x}_2 - \bar{x}_{2s}) \qquad \text{c.f. (3.1)}$$

Note also that $$\hat{\Phi} = I + \hat{B}_{f2} \hat{\Sigma}_{22}^{-1} \hat{B}'_{f2} \qquad \text{c.f. (3.2)}$$

If we finally wish to adopt the normalisation $\Phi = I$ we may take $\hat{\Lambda}\hat{\Phi}^{\frac{1}{2}}$ as the estimated factor loading matrix.

4. Approximate ML Estimator (A)

Suppose that we enter Σ_{11} into a ML factor analysis package. This will give consistent (up to rotation) but inefficient estimators of Λ and Ψ.

5. Approximate ML Estimator (B)

Suppose, as in Method I, we enter S_{11s} into a ML factor analysis package and obtain the estimates $\tilde{\Lambda}$ and $\tilde{\Psi}$ for orthogonal factors $\tilde{f}$. From (4.44) we know that, subject to orthogonal rotations $\tilde{\Lambda} \doteq \Lambda\Phi_s^{\frac{1}{2}}$, $\tilde{f} \doteq \Phi_s^{-\frac{1}{2}} f$. In order to estimate Φ_s we need to estimate B_{f2} from (4.45). Suppose we regress the factors scores $\tilde{f}$ on X_2 as in Joreskog and Goldberger (1975 p.636).

Let $$\tilde{B}_{f2} = \left[\tilde{\Lambda}'(\tilde{\Lambda}\tilde{\Lambda}' + \tilde{\Psi})^{-1}\tilde{\Lambda}\right]^{-1}\tilde{\Lambda}'(\tilde{\Lambda}\tilde{\Lambda}' + \tilde{\Psi})^{-1}\hat{B} \qquad (4.53)$$

where $\hat{B}$ is defined in (3.4)

Then since $B = \Lambda B_{f2} \doteq \tilde{\Lambda}\Phi_s^{-\frac{1}{2}} B_{f2}$ we have

$$\tilde{B}_{f2} \doteq \Phi_s^{-\frac{1}{2}} B_{f2}$$

and from 4.45

$$\Phi_s \doteq I + B_{f2}(S_{22s} - \Sigma_{22})B'_{f2}$$

$$\doteq I + \Phi_s^{\frac{1}{2}}\,\tilde{B}_{f2}(S_{22s} - \Sigma_{22})\tilde{B}'_{f2}\,\Phi_s^{\frac{1}{2}}$$

$$\therefore \quad \Phi_s \doteq (I - \tilde{B}_{f2}(S_{22s} - \Sigma_{22})\tilde{B}'_{f2})^{-1}$$

Hence we might estimate Λ by

$$\tilde{\Lambda}(I - \tilde{B}_{f2}(S_{22s} - \hat{\Sigma}_{22})\tilde{B}'_{f2})^{\frac{1}{2}}$$

and estimate Ψ by $\tilde{\Psi}$

6. Approximate ML Estimator (C)

Jöreskog and Goldberger (1975) consider the same approach as 5 but using the matrix

$$S_{1.2s} = S_{11s} - S_{12s}S_{22s}^{-1}S_{21s}$$

For $n = N = \infty$ the analogous result to (4.44) is

$$S_{1.2s} = \Lambda\Sigma_{f.2}\Lambda' + \Psi$$

Since they adopt the normalisation $\Sigma_{f.2} = I$ it is natural to consider factor analysing $S_{1.2s}$. If we wish to set $\Phi = I$ we must estimate $\Sigma_{f.2}$. Define $\tilde{B}_{f2}$ as in (4.53) where $\tilde{\Lambda}$ and $\tilde{\Psi}$ are obtained from an orthogonal factor analysis of $S_{1.2s}$. Then

$$\tilde{B}_{f2} \doteq \Sigma_{f.2}^{-\frac{1}{2}}B_{f2}$$

Now

$$\Sigma_{f.2} = I - B_{f2}\Sigma_{22}B'_{f2}$$

$$\doteq I - \Sigma_{f.2}^{\frac{1}{2}}\,\tilde{B}_{f2}\,\Sigma_{22}\,\tilde{B}'_{f2}\,\Sigma_{f.2}^{\frac{1}{2}}$$

$\therefore$ an estimate of $\Sigma_{f.2}$ is

$$\hat{\Sigma}_{f.2} = (I - \tilde{B}_{f2}\hat{\Sigma}_{22}\tilde{B}_{f2}')^{-1}$$

and we may estimate Λ by $\tilde{\Lambda}\hat{\Sigma}_{f.2}^{-\frac{1}{2}}$ and Ψ by $\tilde{\Psi}$.

Comparison

The advantages of the exact ML method compared with the approximate ML methods are

(1) It is more efficient.

(2) Standard errors and likelihood ratio tests may be obtained from the LISREL procedure.

Jöreskog and Goldberger (1975) compare the efficiency of the exact ML method in estimating Λ with the approximate method 6. They show theoretically that a substantial gain in efficiency is possible if f is reasonably correlated with X_2. Even so, we suspect that the second reason above is the major advantage of the exact ML method. In the similar problem of factor analysing binary data Muthén (1978) finds that the exact ML method gives very similar estimates to a factor analysis of tetrachoric correlations (c.f. our method 4) and concludes that the main advantage of the former method is that LR tests and standard errors are available.

CHAPTER FIVE - STANDARD ESTIMATORS UNDER TWO-STAGE SAMPLING

5.1 Framework

In Chapters 5, 6 and 7 we consider a two-stage population consisting of N identifiable clusters where the i^{th} cluster contains M_i identifiable units. We consider two possible ways of labelling the units in the finite population.

(i) Lexicographic Labelling: Ordering of the N clusters and of each of the sets of M_i units within each cluster are defined. Then the j^{th} element of the i^{th} cluster is labelled (i,j) for $i=1...N$, $j=1...M_i$.

(ii) Arbitrary Labelling: The $M_o = \Sigma M_i$ units in the population are arbitrarily indexed $k=1...M_o$. Then for each k, x_k is defined as in Example 1.2, i.e. $x_k=e_i$ if k is in the i^{th} cluster (e_i is the $(N-1) \times 1$ vector with 1 in the $(i-1)^{th}$ place and zeros elsewhere, $i=2...N$, $e_1=\underline{0}$). Each unit is then labelled by the pair (k, x_k), $k=1...M_o$.

Let us initially consider the second labelling since it fits in with our notation in Section 1.2. Associated with k^{th} unit of the finite population, U say, is a $p \times 1$ vector y_k. Let $\underline{y} = (y_1'...y_{M_o}')'$, $\underline{x} = (x_1'...x_{M_o}')'$. Then we suppose that a sample, s, is selected from U according to the values $\underline{x}$. As before $\underline{y}$ represents the 'inference' variables and $\underline{x}$ the 'design' variables.

Much of the discussion of this section will be concerned with the choice of an appropriate superpopulation model. As noted in Section 1.3.2., there may be many situations where we are interested in the distribution of $\underline{y}$ given $\underline{x}$. However, in our approach we suppose that the clustering in the population is irrelevant to the target of interest. We therefore wish to define Model I, the 'correct' superpopulation model (see Section 1.2.1), in such a way that we may also define a 'marginal' distribution for 'y' which will be the target of interest. The cluster sizes, M_i, play an important role, so letting $\underline{M} = (M_1...M_N)$, we might express a general superpopulation model as:

$$p(\underline{y}|\underline{x},\underline{M})\ p(\underline{x}|\underline{M})\ p(\underline{M}) \qquad (5.1)$$

Conventional models for $p(\underline{y}|\underline{x}, M)\ p(\underline{x}|\underline{M})$ are given for example by Scott and Smith (1969) and Royall (1976b). In order to obtain the marginal distribution $p(\underline{y})$, however, we need also to specify $p(\underline{M})$. One possibility might be for $\underline{M}$ to be multinomial as in Example 1.3 so that the x_k are IID. This approach seems, however, to be more appropriate for stratification where N is fixed e.g. 2 sexes, 5 social classes etc. A more appropriate model for clusters would seem to us to be where the clusters, e.g. schools or wards, were considered to be a random sample from an infinite population, in which case the M_i will be IID. Such an assumption may also, of course, be necessary for statistical inference when only a subset of the clusters are actually observed.

One problem with such a model is that, although we might be able to integrate $\underline{x}$ out of (5.1), say using the random permutation distribution of Example 1.6, we cannot wholly integrate out $\underline{M}$ because $\underline{y}$ is a $pM_o \times 1$ vector where M_o is a random variable. This problem is no easier under the lexicographic labelling scheme when the labelling of $\underline{y}$ depends not just on M_o but also on $\underline{M}$.

To avoid this difficulty we prefer, instead to make use of the concept, often used by R.A. Fisher, of a superpopulation as an infinite set of units (e.g. Foreman and Brewer, 1971). We think of the finite population as a random sample of clusters from an infinite population of clusters in which the distribution of the values y *per unit* is the target of interest. An example of such a model with a clear generating mechanism is given by Leamer (1978, p.293). Henceforth we use the lexicographic labelling scheme.

Example 5.1

An infinite population (superpopulation) of (completed) families is generated by the following mechanism. Each family continues having children until it has either one boy or two children (if the first child is a girl). The probability of any child being born a boy is λ (uniformly across the population and independent of other births). The clusters consist of the children in each family. The finite

population consists of N such clusters (a random sample of clusters from the superpopulation).

Let M_i = number of children in i^{th} cluster $i = 1...N$

$y_{ij} = 1$ if j^{th} child of i^{th} cluster is a boy $i = 1...N$, $j = 1...M_i$

$= 0$ if j^{th} child of i^{th} cluster is a girl

$$\underline{y} = (y_{11}\cdots y_{1M_1}\cdots y_{NM_N})'$$

For $N = 1$ the p.d.f of $(\underline{y}, \underline{M})$ is

$\underline{y}$	$\underline{M}$	$p(\underline{y},\underline{M})$
(1)	1	λ
(0,1)'	2	$\lambda(1-\lambda)$
(0,0)'	2	$(1-\lambda)^2$

For general N the p.d.f is a product of such densities.

The target of interest in this example is the sex distribution (per child) in the superpopulation. This does not, of course, depend on the families' stopping rule and is

$$P(Y=1) = \lambda \quad , \quad P(Y=0) = 1 - \lambda$$

We refer to this as the 'marginal' distribution of y unconditional on clustering. The question now of interest is : how do we get from $p(\underline{y},\underline{M})$ to this distribution? Let $h_c(M)$ be the proportion of clusters in the superpopulation of size M and let $h_u(M)$ be the proportion of units (children) in the superpopulation which belong to clusters (families) of size M. These are generally not the same and are related by

$$h_u(M) = Mh_c(M)/\Sigma_{M'}\, M'h_c(M') \qquad (5.2)$$

$h_c(M)$ is obtained as the marginal distribution of $\underline{M}$ from $p(\underline{y},\underline{M})$ above.

M	$h_c(M)$	$h_u(M)$
1	λ	$\lambda/(2-\lambda)$
2	$1-\lambda$	$(2-2\lambda)/(2-\lambda)$

The sex distribution in the superpopulation for clusters of a given size is obtained from the conditional distribution, $p(\underline{y}|\underline{M})$.

$\underline{M}$	$\underline{y}$	$p(\underline{y}|\underline{M})$
1	(1)	1
2	(0,1)'	λ
2	(0,0)'	$1-\lambda$

Hence:

Proportion of boys in clusters of size 1 in superpopulation = 1

Proportion of boys in clusters of size 2 in superpopulation = 0.5λ

The required 'marginal' distribution, the sex distribution (per child) in the superpopulation is then obtained as:

$$\text{Proportion of boys in superpopulation} = 1\, h_u(1)+0.5\lambda h_u(2)$$

$$= \lambda/(2-\lambda)+0.5\lambda(2-2\lambda)/(2-\lambda)$$

$$= \lambda$$

Note that this is not the same as

$$\text{average proportion of boys per cluster in superpopulation} = 1h_c(1) + 0.5\lambda h_c(2)$$

$$= \lambda + 0.5\lambda(1-\lambda)$$

$$= 1.5\lambda - 0.5\lambda^2$$

In the light of this example, let us now summarise our approach. The N clusters in the finite population are assumed to be a random sample from an infinite population of clusters. Letting $y_i = (y_{i1}'...y_{iM_i}')'$, we are essentially assuming that the N vectors $(y_i', M_i)'$ $(i = 1...N)$ are IID. Since it is rather non-standard to speak of vectors of different lengths as sharing common distributions we may split our assumption into two parts : (i) $M_1...M_N$ are assumed to be IID, with a common distribution $h_c(M)$, say, (ii) $y_1...y_N$ are independent given $\underline{M}$ where the conditional distribution of y_i given $\underline{M}$ depends only on M_i and is given, say, by $p(y_i|M_i)$. Hence

$$p(\underline{y},\underline{M}) = \prod_{i=1}^{N} p(y_i|M_i)h_c(M_i)$$

The target of interest is conceived of as the 'marginal distribution of y' in the superpopulation. This is obtained from $p(\underline{y},\underline{M})$ as follows. The distribution of the j^{th} cluster member in clusters of size M is defined as

$$f_j(y|M) = \int p\left((y_1'...y_{j-1}'y'y_{j+1}'...y_M')'|M\right) d_{y_1}...d_{y_{j-1}} d_{y_{j+1}}...d_{y_M}$$

The'marginal distribution of y' in clusters of size M is defined as

$$f(y|M) = \sum_{j=1}^{M} f_j(y|M)/M$$

The target of interest, the 'marginal distribution of y' in the superpopulation is defined as

$$p(y) = \sum_{M} f(y|M)h_u(M)$$

where $h_u(M)$ is defined in terms of $h_c(M)$ in (5.2).

Note that, although in Example 5.1 M_i was a 'function' of y_i, the above approach is applicable where y_i and M_i are related in a general manner e.g. if school classes are clusters and y = academic performance then y_i may depend causally on M_i.

The approach above is too general for our purposes, since it takes account of ordering within clusters. We suppose that for practical purposes it is reasonable to assume that the y_{ij} are exchangebly distributed within clusters (e.g. Bellhouse et. al., 1977). In this case $f_j(y|M) = f(y|M)$, $j = 1,..M$. Finally, we make one further assumption. We suppose that the within-cluster distribution of the y_{ij} is IID conditional on a random 'mixing parameter', θ_i. (Note that de Finetti's Theorem states that any *infinite* exchangeable sequence may be so represented, but in our case the M_i are finite). One implication of this assumption is that intracluster correlation must be non-negative which is perhaps undesirable. A more general 'marginal' model as in Royall (1976b) would not be so restrictive (note that our 'mixing' model may be extended as in Walsh (1947) to allow for negative intra-cluster correlations). However, our primary aim is intuitive clarification and we find results for our model are illuminating since we shall be able to distinguish naturally between 'within-cluster' and 'between-cluster' components. The similarity between the results for 'mixing' models and 'marginal' models may be seen by comparing the results of Scott and Smith (1969) and Royall (1976b) for continuous variables and Brier (1980) and Altham (1976) for discrete variables.

Formally, then, we consider a superpopulation model where $(\underline{y},\underline{M})$ is assumed to be a realisation of $(\underline{Y}, \underline{M})$ where $\underline{Y} = (Y_{11}...Y_{NM_N})$ and $\underline{M} = (M_1...M_N)$. For notational covenience we assume that the Y_{ij} are continuous (vector) random variables. We consider two specifications of the joint probability distribution of $(\underline{Y}, \underline{M})$.

Model I (the 'true' Model).

There exist (q-vector) unobserved random variables θ_i, $(i = 1...N)$ such that:

(1) conditional on $\underline{\theta} = (\theta_1...\theta_N)$ and $\underline{M}$ the random variables Y_{ij} $i = 1...N$, $j = 1...M_i$) are independent and Y_{ij} has p.d.f $f(Y|\theta_i, \psi_1)$ indexed by the (vector) parameter ψ_1,

(2) conditional on $\underline{M}$ the random variables θ_i, (i = 1...N), are independent with p.d.f's $g(\theta_i|M_i, \psi_2)$ indexed by the vector parameter ψ_2 (again for notational convenience the θ_i are assumed continuous random variables),

(3) $M_1 \ldots M_N$ are IID random variables with probability mass function $h_c(M_i|\lambda)$ indexed by the (vector) parameter λ.

The joint p.d.f of $(\underline{Y}, \underline{M})$ is thus

$$\prod_{i=1}^{N} \int \left[\prod_{j=1}^{M_i} f(Y_{ij}|\theta_i, \psi_1) \right] g(\theta_i|M_i, \psi_2)\ h_c(M_i|\lambda)\ d\theta_i$$

Model II (the IID Model)

(1) The Y_{ij} are IID given $\underline{M}$ with p.d.f.

$$f_o(Y|\psi_1, \psi_2, \lambda) = \sum_{M} \int f(Y|\theta,\psi_1)\ g(\theta|M,\psi_2) h_u(M|\lambda) d\theta \qquad (5.3)$$

where

$$h_u(M|\lambda) = \frac{M h_c(M|\lambda)}{\sum_{M'} M' h_c(M'|\lambda)}$$

(2) $M_1 \ldots M_N$ are as distributed in Model I (3).
The joint p.d.f. of $(\underline{Y},\underline{M})$ is thus

$$\prod_{i=1}^{N} \left[\prod_{j=1}^{M_i} f_o(Y_{ij}|\psi_1,\psi_2,\lambda) \right] h_c(M_i|\lambda)$$

We now give some examples of our Model I.

Example 5.2

The conventional univariate one-way random effects model (e.g. Novick and Jackson, 1974, p.314) :

$$Y_{ij}|\underline{\theta}, \underline{M} \sim \text{NID}(\theta_i, \psi_1)$$

$$\theta_i|\ \underline{M} \sim \text{NID}(\psi_{21},\psi_{22})$$

is usually expressed conditional on $\underline{M}$. Tan (1978) gives the natural multivariate extension of this model

Example 5.3

Example 5.2 may be extended to allow for dependence of the intra-cluster correlation on $\underline{M}$ according to the usual power-law form (e.g. Hansen et al., 1953, p.307)

$$Y_{ij}|\underline{\theta}, \underline{M} \sim \text{NID}(\theta_i, \psi_1)$$

$$\theta_i|\underline{M} \sim \text{NID}(\psi_{21}, \psi_{22} M_i^{\psi_{23}})$$

Example 5.4

Example 5.2 may also be extended to the heteroskedastic case (e.g. Novick and Jackson, 1974, p.318):

$$Y_{ij}|\underline{\theta}, \underline{M} \sim \text{NID}(\theta_{1i}, \theta_{2i})$$

where θ_{1i} and θ_{2i} are conditionally independent given $\underline{M}$

$$\theta_{1i}|\underline{M} \sim \text{NID}(\psi_{21}, \psi_{22})$$

$$\log\theta_{2i}|\underline{M} \sim \text{NID}(\psi_{23}, \psi_{24})$$

It seems more natural to treat both θ_{1i} and θ_{2i} as random effects rather than treating θ_{1i} as random and θ_{2i} as fixed (e.g. Scott and Smith, 1969; Rao et al., 1981).

Before proceeding we comment again on what is the most unusual feature of our models, the fact that the M_i are taken as random. We have made this assumption in order to define a marginal distribution f_o irrespective of clustering. This assumption was also made by Fuller (1975, Appendix A) in order to facilitate asymptotic arguments. The assumption that the (θ_i, M_i) are jointly IID given (ψ_2, λ) seems to us to be a natural definition of *between-cluster exchangeability*, a definition which has traditionally been problematic. For example,

Rao (1975b) makes the simplifying assumption that θ_i does not depend on M_i and Bellhouse et. al., (1977) attach a rather artificial importance to the quantity $\max(M_i)$. In practice, of course, we shall eventually also need to make restrictive assumptions about the joint distribution of (θ_i, M_i).

We assume that the targets for inference are the parameters of $f_o(Y|\psi_1,\psi_2,\lambda)$ in (5.3). Specifically, as in Chapter 2, we shall be interested in the mean vector, μ, and covariance matrix, Σ, of Y. We now show that under weak conditions the finite population mean, $\bar{Y}$, converges to μ. We may show similarly that the finite population covariance matrix converges to Σ.

Let

$$\bar{Y} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij} / \sum_{i=1}^{N} M_i$$

$$= \sum_{i=1}^{N} M_i \bar{Y}_i \Big/ \sum_{i=1}^{N} M_i$$

where

$$\bar{Y}_i = \sum_{j=1}^{M_i} Y_{ij}/M_i$$

Then

$$E_I(M_i\bar{Y}_i) = \sum_M \iint My f(y|\theta,\psi_1)\, g(\theta|M,\psi_2)\, h_c(M|\lambda)\, d\theta dy$$

$$= \int y f_o(y|\psi_1,\psi_2,\lambda)dy \sum_{M'} M' h_c(M'|\lambda)$$

$$= \mu E(M)$$

Hence, assuming the existence of fourth moments of $M_i\bar{Y}_i$ the Strong Law of Large Numbers implies

$$\Sigma M_i\bar{Y}_i/N \xrightarrow{a.s(I)} \mu E(M)$$

where almost sure convergence under Model I is denoted by a.s(I).

Similarly, assuming the existence of fourth moments of M,

$$\Sigma M_i/N \xrightarrow{a.s(I)} E(M)$$

Hence

$$\bar{Y} \xrightarrow{a.s(I)} \mu$$

Let us now recall our assumptions about the sampling design. A subset s is selected from U. s is a realisation of the random variable S, the distribution of which is defined by the sampling design $p(S|\underline{M} = \underline{M})$. The sampling design is assumed non-informative given $\underline{M} = \underline{M}$, i.e. $\underline{Y}$ and S are conditionally independent given $\underline{M} = \underline{M}$. Without loss of generality let $s = \{(1,1)\ldots(1,m_1)\ldots.(n,m_n)\}$ and write $\underline{y}_s = (y'_{11}\ldots y'_{nm_n})'$ and $\underline{Y}_S$ similarly. The observed data is then $\underline{d} = (\underline{y}_s, s, \underline{M})$ a realisation of $\underline{D} = (\underline{Y}_S, S, \underline{M})$. The p.d.f. of $\underline{D}$ under Model I is

$$p_I(\underline{D}) = \left[\prod_{(i,j)\in S} \int f(Y_{ij}|\theta_i,\psi_1)\, g(\theta_i|M_i,\psi_2)d\theta_i\right] p(S|\underline{M}) \prod_{i=1}^{N} h_c(M_i|\lambda) \qquad (5.4)$$

and under Model II is

$$p_{II}(\underline{D}) = \left[\prod_{(i,j)\in S} f_o(Y_{ij}|\psi_1,\psi_2,\lambda)\right] p(S|\underline{M}) \prod_{i=1}^{N} h_c(M_i|\lambda) \qquad (5.5)$$

Now we should like to be able to make inference conditional on the sample s obtained. Therefore we should like $(s,\underline{M})$ to be ancillary for the parameter of interest (see Section 1.2.3). (Note s always depends on $\underline{M}$ via U). Let us consider two restrictive assumptions about Model I.

Assumption A : (i) θ_i and M_i are independent, $i = 1\ldots N$,

(ii) ψ and λ are Cartesian independent (Definition 1.1), where $\psi = (\psi_1,\psi_2)$.

Assumption B : (i) the distribution of each Y_{ij} gives $\underline{M}=\underline{M}$ does not depend on $\underline{M}$, $i = 1\ldots N$, $j = 1\ldots M_i$, (except in so far as $j \leq M_i$),

(iii) ψ and λ are Cartesian independent.

We shall consider the practical interpretation of these assumptions shortly. We initially discuss their formal implications. Note that Assumption A implies Assumption B, because Y_{ij} only depends on $\underline{M}$ via θ_i. Firstly, we show that under Assumption B (or A) the target of inference is a function of ψ.

Lemma 5.1

Under Assumption B (or A) $f_o(Y|\psi,\lambda)$ does not depend on λ.

Proof:

Under Model I the density of Y_{ij} given $\underline{M}$ is

$$\int f(Y_{ij}|\theta_i,\psi_1)\, g(\theta_i|M_i,\psi_2)d\theta_i = p(Y_{ij}|\psi) ,$$

say, under Assumption B.

Hence from (5.1)

$$f_o(Y|\psi,\lambda) = p(Y|\psi) \sum_{M} h_c(M|\lambda)$$

$$= p(Y|\psi)$$

does not depend on λ.

Now from (5.4) and Definition 1.2. $(s, \underline{M})$ is ancillary for ψ under Model I if Assumption B holds and from (5.5) and Lemma 5.1. $(s, \underline{M})$ is ancillary for ψ under Model II. Hence, according to the Conditionality Principle we may condition on $(s, \underline{M})$ when making inference about ψ if Assumption B(or A) holds.

We now pose two questions : (i) is B likely to hold in practice and (ii) is there a weaker assumption which we might make such that $(s, \underline{M})$ would still be ancillary for the target of inference?

(i) Validity of Assumptions A and B

The important parts of assumptions A and B are contained in the sections (i). Roughly, A makes the assumption that the cluster values y_{ij} are completely independent of cluster size whereas B allows the joint distribution of y_{ij} and $y_{ij'}, (j \neq j')$ to depend on M_i, in particular the intra-cluster correlation may depend on M_i.

It is important to note that these assumptions refer to the *intra-survey* dependence of the y_{ij} on the M_i and not to an *inter-survey* dependence. There have been various empirical investigations of the *inter-survey* dependence of the y's on the M's. For example, Hansen et al (1953 p.588) grouped 1940 census data into equal size clusters of M households by order of enumeration, for M = 3, 9 and 27, and then computed intra-cluster correlations for various variables. This permitted a hypothetical comparison *between* three surveys. On the basis of this and similar evidence Hansen et al state, for example : 'if the units included in the clusters are few and are immediately contiguous there will ordinarily be a higher [intra-cluster] correlation...than when clusters are larger and there is a greater geographic scatter of the units' (p.262).

Such results are important in considerations of the optimal design of two-stage surveys (e.g. Hansen et al., 1953, p.306; Cochran, 1977, p.244; Brewer et al., 1977). However, the same empirical evidence has been used to justify models used for *intra-survey* comparison of estimators (e.g. Des Raj, 1958; Rao, 1967; Foreman and Brewer, 1971; Royall, 1976b; Cochran, 1977, p.256). We use an example to emphasize the distinction. Consider two surveys of the same population, one of which uses census enumeration districts (ED's) as clusters and the other of which uses wards (comprising several ED's). In accordance with Hansen et al's statement on inter-survey comparisons it would appear almost axiomatic that the average intracluster correlation for the first survey would be not less than the average intracluster correlation in the second for any variable. However, consider just the first survey. In order to equalize workloads, ED's in homogenous areas are relatively large and ED's in 'difficult' areas, such as those containing multi-occupancy households, are relatively small. Hence for many variables

the intra-cluster correlation may be higher for the larger clusters than for the smaller clusters in contrast with the inter-survey situation.

In general it is not difficult to conceive of situations where Assumption A and B might fail to hold. For example, in national surveys with clusters as geographical regions M_i will usually be larger in urban clusters than in rural clusters and many y variables may be correlated with the urban/rural dichotomy. However, such differences are usually allowed for by stratification and Assumption B may be reasonable within strata. Some diagnostic checks of assumptions A and B are given in Sections 5.3 - 5.5 where, for a particular data set on schools, both assumptions seem plausible.

Even under these assumptions our model is still as general as most models in the literature (e.g. Rao and Scott's (1981) extension of Brier's (1980) model to the case of unequal M_i's essentially makes Assumption A). If Assumption B does not hold then an approach as in Chapter 3 might be possible setting $x_{1i} = y_i$, $x_{2i} = M_i$.

(ii) A necessary condition for $(s, \underline{M})$ to be ancillary for the target of inference.

We know that Assumption B is a sufficient condition for $(s, \underline{M})$ to be ancillary for the target of inference. A necessary and sufficient condition is that

(i) $f_o(Y|\psi,\lambda)$ does not depend on λ

(ii) ψ and λ are Cartesian independent.

We give two counter-examples to show that B is not a necessary condition.

(a) If λ is known then $f_o(Y|\psi,\lambda)$ trivially does not depend on λ although Y_{ij} and M_i may still be dependent.

(b) Let λ take two values, λ_1 and λ_2.

Let

$$h(M|\lambda) = \frac{1}{2}\ ,\ M = 1,\ 3 \quad \text{if} \quad \lambda = \lambda_1$$

$$= \frac{1}{2}\ ,\ M = 2,\ 4 \quad \text{if} \quad \lambda = \lambda_2$$

Let

$$\int f(Y|\theta,\psi_1)\ g(\theta|M,\psi_2)d\ = f_1(Y|\psi) \quad \text{if} \quad M = 1 \quad \text{or} \quad 2$$

$$f_2(Y|\psi) \quad \text{if} \quad M = 3 \quad \text{or} \quad 4$$

Then

$$f_o(Y|\psi,\lambda) = \left[f_1(Y|\psi) + f_2(Y|\psi)\right]/2 \quad \text{if} \quad = \lambda_1 \ \underline{\text{or}}\ \lambda_2$$

Hence f_o does not depend on λ but Y and M are not independent.

Such examples are, however, rather artificial and do not seem to possess any natural practical interpretation. We therefore feel that B is an adequately general sufficient condition for inference to be conditional on $(s, \underline{M})$.

For the remainder of this chapter we propose to evaluate the properties of standard estimators (i.e. estimators under the assumption that Model II is correct) under Model I subject to Assumptions A and B and under Model II. Given the argument above we shall evaluate the sampling distributions of estimators conditional on $(s, \underline{M})$.

5.2 General Properties of Standard Estimators

The most general form of a standard estimator (i.e. an estimator based on the assumption that Model II is correct) that we shall consider is $g(\underline{T})$, where g is a given real-valued function, $\underline{T} = (T_1 \ldots T_p)$,

$$T_h = \sum_{(i,j)\epsilon s} h(y_{ij}) \quad , \quad h = 1 \ldots p, \qquad (5.6)$$

and h is a given real-valued function. This definition includes means, variances, covariances, correlation coefficients and regression coefficients (c.f.Krew ski and Rao, 1981). For example, if $h(y) = y/(m_1 + \ldots m_n)$ then T_h is the sample mean. Note that in general h may depend on the quantities $m_1 \ldots m_n$ and n which might vary between samples, s.

We initially define some notation.

Let

$$s_1 = \{i;\ (i,j) \in s \quad \text{for some} \quad j\}$$

$$s_{2i} = \{j;\ (i,j) \in s\}$$

Without loss of generality suppose

$$s_1 = \{1 \ldots n\} \quad , \quad s_{2i} = \{1 \ldots m_i\}$$

Let

$$m_o = \sum_{i=1}^{n} m_i \quad , \qquad m^* = \sum_{i=1}^{n} m_i^2/m_o \tag{5.7}$$

Let μ_h and σ_h^2 be the mean and variance of $h(y)$ in the superpopulation i.e.

$$\mu_h = \int h(y) f_o(y|\psi,\lambda)dy \tag{5.8}$$

$$\sigma_h^2 = \int \left(h(y) - \mu_h\right)^2 f_o(y|\psi,\lambda)dy \tag{5.9}$$

where f_o is defined in (5.3).

Let $\tau_h(M)$ be the intra-cluster correlation of $h(y)$ in a cluster of size M under Model I, i.e.

$$\tau_h(M) = \text{corr}_I\left[h(Y_{ij}),\ h(Y_{ij'})\,|\,M_i = M\right] \qquad j \neq j' \tag{5.10}$$

We now consider the properties of the estimator, T_h, and then consider the more general estimator, $g(\underline{T})$.

Lemma 5.2

If Assumption B holds

$$E_I(T_h|s,\underline{M}) = E_{II}(T_h|s,\underline{M}) = m_o\mu_h$$

$$V_I(T_h|s,\underline{M}) = \left[1 + \sum_{i=1}^{n} m_i(m_i-1)\,\tau_h(M_i)/m_o\right] m_o\sigma_h^2$$

$$V_{II}(T_h|s,\underline{M}) = m_o\sigma_h^2$$

Proof:

From Lemma 5.1, if B holds the distribution of Y_{ij} given $\underline{M} = \underline{M}$ under Model I is $f_o(Y_{ij}|\psi)$. Hence

$$E_I(T_h|s,\underline{M}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} h(Y_{ij}) f_o(Y_{ij}|\psi) dY_{ij}$$

$$= m_o \mu_h$$

$$= E_{II}(T_h|s,\underline{M})$$

$$V_I(T_h|s,\underline{M}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} V_I\left[h(Y_{ij})|M_i\right]$$

$$+ \sum_{i=1}^{n} \sum_{j \neq j'} \mathrm{cov}_I\left[h(Y_{ij}), h(Y_{ij'})|M_i\right]$$

$$= m_o \sigma_h^2 + \Sigma m_i(m_i-1)\tau_h(M_i)\sigma_h^2 \quad \text{if B holds}$$

$$= \left[1 + \Sigma m_i(m_i-1)\tau_h(M_i)/m_o\right] m_o \sigma_h^2$$

$$V_{II}(T_h|s,\underline{M}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} V_{II}\left(h(Y_{ij})|\underline{M}\right)$$

$$= m_o \sigma_h^2$$

Corollary 5.2

If Assumption A holds

$$V_I(T_h|s,\underline{M}) = \left[1 + (m^*-1)\tau_h\right] m_o \sigma_h^2$$

where $\tau_h(M) = \tau_h$ does not depend on M. We shall also require the covariances between the T_h and T_k . The analogous results are given in the following Lemma.

Lemma 5.3

If Assumption B holds then

$$\mathrm{cov}_I(T_h, T_k|s,\underline{M}) = \left[1 + \sum_{i=1}^{n} m_i(m_i-1)\tau_{hk}(M_i)/m_o\right] m_o \sigma_{hk}$$

where

$$\tau_{hk}(M) = \text{cov}_I\left[h(Y_{ij}),\ k(Y_{ij'}) \middle| M_i = M_i\right]/\sigma_{hk}, \quad j \neq j' \tag{5.11}$$

$$\sigma_{hk} = \int \big(h(y) - \mu_h\big)\big(k(y) - \mu_k\big) f_o(y|\psi,\lambda)\,dy \tag{5.12}$$

Proof:

If B holds then the distribution of Y_{ij} given $\underline{M} = \underline{M}$ under Model I is $f_o(Y_{ij}|\psi)$. Hence

$$\text{cov}_I\left[h(Y_{ij}),\ k(Y_{ij}) \middle| \underline{M} = \underline{M}\right] = \sigma_{hk}$$

Now

$$\text{cov}_I(T_h,\ T_k|s,\underline{M}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \text{cov}_I\left[h(Y_{ij}),\ k(Y_{ij}) \middle| \underline{M}\right]$$

$$+ \sum_{i=1}^{n} \sum_{j\neq j'} \text{cov}_I\left[h(Y_{ij}),\ k(Y_{ij'}) \middle| \underline{M}\right]$$

$$= \left[1 + \Sigma m_i(m_i-1)\tau_{hk}(M_i)/m_o\right] m_o\sigma_{hk}$$

as required.

To obtain corresponding results for $g(\underline{T})$ we use a Taylor Series Linearisation. For this purpose we need to adopt a limiting argument. It is more straightforward to index the limits by n (as in Fuller, 1975) that by N as in Section 2.1. We therefore consider a nested sequence of finite populations, U_n , containing N_n clusters, where each of the sets of finite population values $(\underline{y}, \underline{M})$ obey Model I. In addition we define a sequence of sampling designs $p_n(s|\underline{M})$.

Let F_n be the distribution of a statistic $T = T(\underline{y}_s)$ with respect to Model I conditional on $(s_n, \underline{M}_n)$. We shall say that

$$T \xrightarrow{L} F \quad \text{a.s.}$$

if

$$P\big(F_n(t) \to F(t) \quad \text{for all} \quad t\ \big) = 1$$

where the probability measure is taken with respect to the joint distribution of $(s_n, \underline{M}_n)$.

We shall require the following three conditions.

$$\text{C1: } \left(\sum_{i=1}^{n} m_i\right)\mu_h \xrightarrow{a.s} \tilde{\mu}_h \qquad h = 1\ldots p$$

$$\text{C2: } n \sum_{i=1}^{n} m_i\left[1 + (m_i-1)\tau_{hk}(M_i)\right]\sigma_{hk} \xrightarrow{a.s} \tilde{\sigma}_{hk} \quad h,k = 1\ldots p$$

$$\text{C3: } n \sum_{i=1}^{n} E\left|\alpha_{hi} - nm_i\mu_h\right|^{2+\delta} \text{ is bounded a.s } \quad h = 1\ldots p \text{ for some } \delta > 0$$

where

$$\alpha_{hi} = \sum_{j=1}^{m_i} h(y_{ij})$$

These conditions seem reasonable for the h types functions that we shall consider. Generally h will either involve a denominator $(\Sigma m_i)^{-1}$ or $(\Sigma m_i - 1)^{-1}$ and hence condition C1 essentially just requires the existence of the moments of Y_{ij}. Condition C2 involves a similar requirement in addition to the coefficient of variation of the M_i being bounded and the τ_h functions being regular. Condition C3 is a standard Liapounov-type condition.

Lemma 5.4

If Assumption B and Conditions C1, C2 and C3 hold then under Model I

$$\sqrt{n}\left[\begin{pmatrix} T_1 \\ \vdots \\ T_p \end{pmatrix} - \begin{pmatrix} \tilde{\mu}_1 \\ \vdots \\ \tilde{\mu}_p \end{pmatrix}\right] \xrightarrow{L} N_p(0, \tilde{\Sigma}) \quad a.s$$

where

$$\tilde{\Sigma}_{hk} = \tilde{\sigma}_{hk}$$

Proof :

Write

$$T_h = \sum_{i=1}^{n} t_{hi}/n$$

where

$$t_{hi} = n\alpha_{hi} = n \sum_{j=1}^{m_i} h(Y_{ij})$$

then

$$E_I(t_{hi}|s,\underline{M}) = nm_i\mu_h$$

$$cov_I(t_{hi}, t_{ki}|s,\underline{M}) = n^2m_i\left(1+(m_i-1)\tau_{hk}(M_i)\right)\sigma_{hk}$$

from Lemma 5.3.

The result then follows by applying a Central Limit Theorem for independent non-identically distributed random variables (e.g. Krewski and Rao, 1981, Lemma 3.1).

In order to obtain the corresponding result for $g(\underline{T})$ we need one more condition : C4 : g is a real-valued continuous function with continuous first and second derivatives at the point $\underline{\tilde{\mu}} = (\tilde{\mu}_1 \ldots \tilde{\mu}_p)$.

Lemma 5.5

If Assumption B and conditions C1 - C4 hold then under Model I

$$\sqrt{n}\left[g(\underline{T}) - g(\underline{\tilde{\mu}})\right] \xrightarrow{L} N(0, \sum_{hk} g_h(\underline{\tilde{\mu}})g_k(\underline{\tilde{\mu}})\tilde{\sigma}_{hk}) \quad a.s$$

where $g_h(\underline{\tilde{\mu}})$ is the partial derivative of $g(\underline{T})$ with respect to T_h evaluated at $\underline{T} = \underline{\tilde{\mu}}$.

Proof :

This follows from Lemma 5.4 by standard Taylor Series Linearisation (e.g. Rao, 1973, p.387).

Corollary 5.5

We may write

$$V_I\left[g(\underline{T})|s,\underline{M}\right] \doteq \left[1 + \Sigma m_i(m_i-1)\tau_g(M_i)/m_o\right]m_o\Sigma\Sigma\beta_{k\ell}$$

$$V_{II}\left[g(\underline{T})|s,\underline{M}\right] \doteq m_o\Sigma\Sigma\beta_{k\ell}$$

where

$$\tau_g(M_i) = \sum_k \sum_\ell \alpha_{k\ell}\, \tau_{k\ell}(M_i) \qquad (5.13)$$

$$\alpha_{k\ell} = \beta_{k\ell} / \Sigma\Sigma\beta_{k\ell}$$

$$\beta_{k\ell} = g_k(\tilde{\mu}) g_\ell(\tilde{\mu}) \sigma_{k\ell}$$

Proof:

From Lemma 5.5.

$$V_I(g(\underline{T})|s,\underline{M}) = \sum_{hk} g_h(\tilde{\mu}) g_k(\tilde{\mu}) \tilde{\sigma}_{hk}/n$$

$$\doteq \sum_{hk} g_h(\tilde{\mu}) g_k(\tilde{\mu}) \sum_{i=1}^{n} m_i \left[1 + (m_i - 1)\tau_{hk}(M_i)\right] \sigma_{hk}$$

from C2

$$= m_o \Sigma\Sigma\beta_{hk} + \sum_{i=1}^{n} m_i(m_i - 1) \sum_{hk} \beta_{hk} \tau_{hk}(M_i)$$

$$= \left[1 + \Sigma m_i(m_i - 1)\tau_g(M_i)\right] m_o \sum_{hk} \beta_{hk}$$

as required.

Let us now consider the results of Lemmas 5.2 and 5.5. Under Assumption B we see that misspecifying the model as Model II instead of Model I only affects the variance of T_h and does not introduce bias. For $g(\underline{T})$ there may be a 'misspecification bias' but this of an order smaller than the effect of misspecification on the variance. This is in direct contrast with the results of Chapter 2 (e.g. Theorems 2.1, 2.4 and 2.10) where the main effect of misspecification was in terms of bias rather than variance. These results are, however, in accordance with the empirical work of Kish and Frankel (1974) who generally found negligible design effects on bias(except for the multiple correlation coefficient).

On the basis of these results we propose in the remainder of this chapter (and also in Chapter 7) to summarise the effect of misspecification by a single quantity.

Definition 5.1 : If Assumption B holds and T is a 'standard' estimator of ψ then the *(conditional) misspecification effect* of T is defined as

$$\text{meff}(T|s,\underline{M}) = V_I(T|s,\underline{M})/V_{II}(T|s,\underline{M}) \tag{5.14}$$

Hence from Lemma 5.2.

$$\text{meff}(T_h|s,\underline{M}) = 1 + \Sigma\, m_i(m_i-1)\tau_h(M_i)/m_o \tag{5.15}$$

and from Corollary 5.5

$$\text{meff}(g(\underline{T})|s,\underline{M}) \doteq 1 + \Sigma m_i(m_i-1)\tau_g(M_i)/m_o \tag{5.16}$$

If Assumption A holds then from Corollary 5.2

$$\text{meff}(T_h|s,\underline{M}) = 1 + (m^*-1)\tau_h \tag{5.17}$$

and similarly

$$\text{meff}(g(\underline{T})|s,\underline{M}) \doteq 1 + (m^*-1)\tau_g \tag{5.18}$$

If the m_i are all equal to m then $m^* = m$ and the misspecification effect has the familiar form of a design effect for cluster sampling where τ is the intra-cluster correlation (e.g. Kish, 1965, 8.2). Let us consider this analogy more formally.

Definition 5.2: Suppose $p(s|\underline{M})$ is a fixed size design of size $n = n(\underline{M})$. The *design effect* of T is defined as

$$\text{deff}(T|\underline{y},\underline{M}) = V_p(T|\underline{y},\underline{M})/V_{p_o}(T|\underline{y},\underline{M})$$

where

$$V_p(T|\underline{y},\underline{M}) = \sum_s (T(\underline{y}_s) - \bar{T})^2 p(s|\underline{M})$$

$$\bar{T} = \sum_s T(\underline{y}_s)\, p(s|\underline{M})$$

$$V_{p_o}(T|\underline{y},\underline{M}) = \sum_s (T(\underline{y}_s) - \bar{T}_o)^2\, p_o(s|\underline{M})$$

$$\bar{T}_o = \sum_s T(y_s)\, p_o(s|\underline{M})$$

$P_o(s|\underline{M})$ is the (simple random) sampling design which assigns equal probability to all possible samples of size $n(\underline{M})$ from U, and the notation $T(\underline{y}_s)$ is used to emphasise that T depends on s.

We now show why our results for misspecification effects correspond to formulae for design effects when $m_i = m$.

Lemma 5.6

Let $p_1(S|\underline{M})$ denote a with-replacement PPS epsem design, i.e. let $S_{n,N}$ denote the set of ordered subsets of {1 ... N} of size n.

Let $S_1 = \left\{ s \; ; \; s = \{(i,j) \; ; \; i \in S_{n,N}, \; j \in S_{m,M_i}\} \right\}$

Write $i \in s$ if $(i,j) \in s$ for some $j \in S_{m,M_i}$.

Then
$$p_1(s|\underline{M}) = \prod_{i \in s} \frac{M_i}{M_o} \left(\frac{1}{M_i}\right)^m \qquad s \in S_1$$
$$= 0 \qquad s \notin S_1$$

Let $p_2(S|\underline{M})$ derote the srswr design with same fixed size as p_1 i.e. let S_2 denote the set of all samples of size nm from $U(\underline{M})$.

Then
$$p_2(s|\underline{\zeta}=\underline{M}) = \left(\frac{1}{M_o}\right)^{nm} \qquad \text{if } s \in S_2$$
$$= 0 \qquad \text{if } s \notin S_2$$

Then the sampling distribution of y_s under p_1 obeys assumption (1) and (2) of Model I with $(M_i, y_{i1} \ldots y_{iMi})$ substituted for θ_i and $(y_{11} \ldots y_{NM_N})$ substituted for ψ. Similarly the sampling distribution of $\underline{y_s}$ under p_2 obeys assumption (i) of Model II with the same substituted value for ψ. Furthermore the marginal distribution of y_{ij}, which is the same under p_1 and p_2, does not depend on $\underline{M}$ and hence, under our analogy, Assumption B holds.

<u>Proof</u>: Let $\theta_i = (M_i, y_{i1} \cdots y_{iM_i})$ $\quad i=1 \ldots N$

Then (1) conditional on θ_i ($i \in s$) the components y_{ij} of $\underline{y}_s$ are independent under p_1 with

$$P(y_{ij} = y_{k\ell} | \theta_i = \theta_t) = 1/M_t \quad \text{if } t=k \quad \text{for } \ell=1 \ldots M_t$$

$$= 0 \quad \text{if } t \neq k$$

and (2) conditional on $\underline{M}$, θ_i($i \in s$) are independent with

$$P(\theta_i = \theta_t | \underline{M}) = M_t/M_o \quad t=1 \ldots N$$

The marginal distribution of y_{ij} ($i \in s$) is then

$$P(y_{ij} = y_{k\ell} | \underline{M}, \psi) = \frac{1}{M_k} \cdot \frac{M_k}{M_o} = 1/M_o \quad k=1 \ldots N, \ \ell=1 \ldots M_k$$

For p_2 the components y_{ij} of $\underline{y}_s$ are iid with

$$P(y_{ij} = y_{k\ell} | \underline{M}, \psi) = 1/M_o \quad k = 1 \ldots N, \ \ell = 1 \ldots M_k$$

<u>Corollary 5.7</u>

Formulae for design effects with respect to the designs p_1 and p_2 of Lemma 5.6 may be obtained as particular cases of general formulae for (conditional) misspecification effects using Models I and II.

<u>Lemma 5.8</u>

The only designs which are isomorphic to Models I and II in the sense of Lemma 5.6 and for which B holds are given by p_1 and p_2 in Lemma 5.6.

<u>Proof</u>

In order for there to be independence within and between clusters in Model I, p_1 must be of the form:

$$p_1(s|\underline{M}) = \prod_{i \in s} a(M_i)\left(\frac{1}{M_i}\right)^{b(M_i)}$$

where $a(M_i)$ and $b(M_i)$ are functions of M_i. The total sample size is then

$$m_o = \sum_{i \varepsilon s} b(M_i)$$

But this is the length of the vector $\underline{y_s}$ which has fixed length under Model I. Hence m_o must be fixed under p_1 and, since the clusters are selected independently with replacement, $b(M_i)$ must be a constant, $b(M_i) = m$. Similarly the number of clusters n must be a constant.

Now the marginal distribution of a component y_{ij} of $\underline{y_s}$ under p_1 is given by

$$P(y_{ij} = y_{k\ell} | \underline{M}) = a(M_k)/M_k \qquad k=1 \ldots N$$

If B holds this must not depend on M_k and hence p_1 must be given by the PPS design in Lemma 5.6 Finally since the marginal distribution of y_{ij} must be the same under p_1 and p_2 we must have p_2 as given in Lemma 5.6

To summarise, formulae that we shall obtain for misspecification effects will also be applicable to design effects under the design in Lemma 5.6 subject to reparametrisation. Note that design effects are often used for inferential purposes as proportional adjustments for simple random sampling variance formulae and in this sense are not dissimilar in interpretation to misspecification effects.

Let us now consider more closely the formulae for misspecification effects. These depend on the population structure through $\tau_h(M)$ and $\tau_g(M)$ and on the sample, s, selected through the m_i. If Assumption A holds then the form $1 + (m^* - 1)\tau$ has also been derived by Campbell (1977), Holt (1980) and Rao and Scott (1981). Note that this differs from the formula $1 + (\bar{m}-1)roh$ used e.g. by Kish et al. (1976). If we equate these terms

$$1 + (m^*-1)\tau = 1 + (\bar{m} - 1)roh$$

then

$$|roh| = \frac{m^*-1}{\bar{m}-1} |\tau|$$

$$\geq |\tau|$$

Since

$$m^* = \bar{m} + \Sigma(m_i - \bar{m})^2/m_o \geq \bar{m}$$

Hence Kish et al's (1976) roh will in general be greater than our τ.

From (5.10), $\tau_h(M)$ may be interpreted as a generalised intra-cluster correlation. In the following lemma we show that the same is true for $\tau_g(M)$ defined in (5.13) (c.f. Woodruff, 1971).

Lemma 5.9

If B holds

$$\tau_g(M_i) = \text{corr}_I(Z_{ij}, Z_{ij'} | M_i) \quad j \neq j'$$

where

$$Z_{ij} = \sum_h h(Y_{ij})\ g_h(m_o\underline{\mu})$$

Proof:

From (5.13) and (5.11)

$$\tau_g(M_i) = \sum_h \sum_k g_h(m_o\underline{\mu}) g_k(m_o\underline{\mu}) \text{cov}_I\left[h(Y_{ij}), k(Y_{ij'}) | M_i\right] / \sum_h \sum_k g_h(m_o\underline{\mu}) g_k(m_o\underline{\mu}) \sigma_{hk}$$
$$= \text{cov}_I\left[Z_{ij}, Z_{ij'} | M_i\right] / \text{var}_I\left[Z_{ij}\right]$$

as required.

All the results obtained so far in this section have only been based on the within-cluster exchangeability of the Y_{ij}. Using the θ_i we may now obtain another representation of the τ's.

Let
$$\mu_{hi} = E_I\left(h(Y_{ij}) | \theta_i\right) \tag{5.19}$$

Then if B holds

$$\tau_h(M_i) = V_I(\mu_{hi} | M_i)/\sigma_h^2 \tag{5.20}$$

i.e. τ_h measures the between-cluster variation in $h(Y_{ij})$ relative to the overall variation. Note that σ_h^2 may be decomposed into between and within-cluster components

$$\sigma_h^2 = V_I(\mu_{hi}|M_i) + E_I(\sigma_{hi}^2|M_i) \tag{5.21}$$

where
$$\sigma_{hi}^2 = V_I\left(h(Y_{ij})|\theta_i\right) \tag{5.22}$$

Analogous approximate results are also available for τ_g.

Lemma 5.10

If B holds

$$\tau_g(M_i) \doteq V_I\left[g(\underline{U}_i)|M_i\right]/m_o V_{II}\left[g(\underline{T})|s,\underline{M}\right]$$

where
$$U_i = (U_{i1}\cdots U_{ip})$$

$$U_{ih} = m_o\mu_{hi}$$

Proof:

For Z_{ij} defined in Lemma 5.9

$$E_I(Z_{ij}|\theta_i) = \Sigma\, g_h(m_o\underline{\mu})\mu_{hi}$$

Hence from Lemma 5.9

$$\tau_g(M_i) = \frac{V_I\left(\sum_h g_h(m_o\underline{\mu})\mu_{hi}|M_i\right)}{V_I\left(\sum_h g_h(m_o\underline{\mu})h(Y_{ij})\right)}$$

$$= \frac{V_I\left(\sum_h g_h(m_o\underline{\mu})m_o\mu_{hi}|M_i\right)}{m_o V_{II}\left(\sum_h g_h(m_o\underline{\mu})T_h|s,\underline{M}\right)}$$

Now

$$E_I(m_o\mu_{hi}|M_i) = m_o\mu_h$$

Hence

$$\tau_g(M_i) \doteq \frac{V_I\left(g(\underline{U}_i)\,|\,M_i\right)}{m_o V_{II}(g(\underline{T})\,|\,s,\underline{M})}$$

as required.

Note that this result does not involve partial derivatives and we suggest that it might have uses in variance estimation.

In this section we have argued that given Assumption B for our Models I and II the main effect of misspecification may be summarised in a single measure. This 'misspecification effect' depends in the same way on the design for any statistic $g(\underline{T})$. The only difference occurs in the intracluster correlations $\tau_g(M)$. In the remainder of this chapter we shall be concerned with the form of $\tau_g(M)$ for various statistics $g(\underline{T})$. Finally in this section, however, we consider the implications of misspecification when Assumption B does not hold.

If B does not hold then $f_o(Y|\psi,\lambda)$ (defined in 5.3)) depends in general on λ as well as ψ. The target of inference is therefore generally a function of λ and ψ and as argued in Section 5.1 $(s,\underline{M})$ is no longer ancillary for this target of inference. Hence we consider properties of estimators unconditional on s and $\underline{M}$. To see how different this situation can be we initially take an example.

Example 5.5

Let us adapt Example 5.1. Suppose $M_i = 1$ or 2 with $h_c(1) = \lambda$, $h_c(2) = 1 - \lambda$. Suppose that $Y_{ij} = 1$ with probability one in clusters of size 1 and suppose that in clusters of size 2

$$P\left[(Y_{i1},Y_{i2}) = (0,\ 0)\,|\,M_i = 2\right] = 1 - \lambda$$

$$P\left[(Y_{i1},Y_{i2}) = (0,\ 1)\,|\,M_i = 2\right] = P\left[(Y_{i1},Y_{i2}) = (1,\ 0)\,|\,M_i\right] = \lambda/2$$

Suppose that $N = \infty$ and that we select just one cluster of size M with probability α for $M = 1$ and $1 - \alpha$ for $M = 2$. If $M = 2$ suppose we select one unit from the cluster at random. Let T be the single observed value y_{ij}. Then

$$E_I(T) = \alpha + (1-\alpha)\lambda/2$$

$$E_{II}(T) = \lambda$$

$$V_I(T) = (1-\alpha)\lambda/2(1-\lambda/2) + \alpha + (1-\alpha)\lambda^2/4 - (\alpha+(1-\alpha)\lambda/2)^2$$

$$V_{II}(T) = \lambda(1-\lambda)$$

Hence there is a misspecification bias unless $\alpha + (1-\alpha)\lambda/2 = \lambda$ i.e. if $\alpha = \lambda/(2-\lambda)$ i.e. if the sampling is PPS. Note also that the 'misspecification effect' above, $V_I(T)/V_{II}(T)$, is unbounded since as $\lambda \to 0$, $V_I(T) \to \alpha(1-\alpha)$ and $V_{II}(T) \to 0$. More practically, suppose we select n such values independently. Then the misspecification bias of the sample mean is exactly as above wheareas the variances under both Models are of $O(n^{-1})$. This situation is more akin to Chapter 2 than to other results in this section.

One situation where the $p\xi$ - misspecification effect will only act via the variance rather than bias is when the design is self-weighting. We now consider such effects for two self-weighting designs. For simplicity we restrict attention to the sample mean

$$\bar{y}_s = \sum_{i=1}^{n} \sum_{j=1}^{\dot{m}_i} y_{ij}/m_o$$

where y_{ij} is assumed univariate. Note that distributions will be taken unconditionally and the limits are evaluated as discussed after Corollary 5.3. Note also that finite moment assumptions as in Fuller (1975, Theorem A) are assumed.

Lemma 5.11a

If B does not hold and the sampling design is the self-weighting PPS design given by p_1 in Lemma 5.6 with fixed m then

$$E_I(\bar{y}_s) \to \mu \qquad \text{as} \quad N \to \infty$$

$$E_{II}(\bar{y}_s) = \mu$$

$$nV_I(\bar{y}_s) \to \left[1 + (m-1)\tau\right]\sigma^2/m \qquad \text{as} \quad N \to \infty$$

$$nV_{II}(\bar{y}_s) = \sigma^2/m$$

where μ and σ^2 are the mean and variance of Y_{ij} in the distribution f_o of (5.3) and

$$\tau = \text{corr}_I(Y_{ij}, Y'_{ij}) \qquad j \neq j'$$

is evaluated with respect to the distribution $h_u(M)$ defined in (5.2)

Proof:

Let $f(y|M)$ be the distribution of Y_{ij} given M_i

i.e. $$f(y|M) = \int f(y|\theta, \psi_1)\, g(\theta|M, \psi_2)\, d\theta$$

Let $$\mu(M) = \int y f(y|M)\, dy$$

Then $$E_I(\bar{y}_s) = E_I\left\{E_I\left[E_I(\bar{y}_s|s, \underline{M})|\underline{M}\right]\right\}$$

$$= E_I\left\{E_I\left[\sum_{i=1}^{n} m\,\mu(M_i)/nm|\underline{M}\right]\right\}$$

$$= E_I\left\{\frac{1}{N}\sum_{i=1}^{N} M_i\mu(M_i)/\frac{1}{N}\sum_{i=1}^{N} M_i\right\}$$

$$\to \frac{\Sigma M\mu(M)h_c(M)}{\Sigma M h_c(M)} \qquad \text{as} \quad N \to \infty$$

$$= \Sigma\mu(M)h_u(M)$$

$$= \mu$$

$$E_{II}(y_s) = \mu \text{ as in Lemma 5.2.}$$

Let $$\sigma_B^2(M_i) = V_I(\mu_i|M_i)$$

$$\sigma_W^2(M_i) = E_I(\sigma_i^2|M_i)$$

where $\mu_i = E_I(Y_{ij}|\theta_i)$, $\sigma_i^2 = V_I(Y_{ij}|\theta_i)$ as in (5.19) and (5.22).

Then

$$V_I(\bar{y}_s) = V_I\{E_I(\bar{y}_s|s, \underline{M})\} + E_I\{V_I(\bar{y}_s|s, \underline{M})\}$$

$$= V_I\left\{\sum_{i=1}^{n} \mu(M_i)/n\right\} + E_I\left\{\sum_{i=1}^{n} \sigma_B^2(M_i)/n^2 + \sum_{i=1}^{n} \sigma_W^2(M_i)/n^2m\right\}$$

$$= V_I\left\{\sum_{i=1}^{N} M_i\mu(M_i)/\sum_{i=1}^{N} M_i\right\} + E_I\left\{\sum_{i=1}^{N} M_i(\mu(M_i)-\bar{\mu})^2/\sum_{i=1}^{N} M_i\right\}/n$$

$$+ E_I\left\{\sum_{i=1}^{N} M_i(m\sigma_B^2(M_i) + \sigma_W^2(M_i))/\sum_{i=1}^{N} M_i\right\}/nm$$

where

$$\bar{\mu} = \sum_{i=1}^{N} M_i\mu(M_i)/\sum_{i=1}^{N} M_i$$

$$\therefore \quad nV_I(\bar{y}_s) \to \Sigma M(\mu(M)-\mu)^2\, h_c(M)/\Sigma h_c(M) + \Sigma M(m\sigma_B^2(M)+\sigma_W^2(M))h_c(M)/\Sigma h_c(M)m$$

as $N \to \infty$

provided $n/N \to 0$ as $N \to \infty$

$$\therefore \quad nV_I(\bar{y}_s) \to \Sigma(\mu(M)-\mu)^2 h_u(M)$$

$$+ \Sigma(m\sigma_B^2(M)+\sigma_W^2(M))h_u(M)/m \qquad (5.23)$$

Now

$$\tau = \text{corr}_I(Y_{ij}, Y_{ij'}) \qquad j \neq j$$

$$= V_I(\mu_i)/\sigma^2$$

$$= \left[\Sigma\sigma_B^2(M)h_u(M) + \Sigma(\mu(M)-\mu)^2h_u(M)\right]/\sigma^2 \qquad (5.24)$$

$$\sigma^2 = \Sigma\sigma_W^2(M)h_u(M) + \Sigma\sigma_B^2(M)h_u(M) + \Sigma(\mu(M)-\mu)^2h_u(M) \qquad (5.25)$$

Combining (5.23) - (5.25)

$$nV_I(\bar{y}_s) \to (1 + (m-1)\tau)\sigma^2/m \quad \text{as required}$$

Finally $nV_{II}(\bar{y}_s) = \sigma^2/m$ follows from Lemma 5.2.

Lemma 5.11b

If B does not hold and the sampling design is a single-stage clustered design (where the clusters are selected by srswor) then

$$E_I(\bar{y}_s) \to \mu \qquad \text{as} \quad N \to \infty$$

$$E_{II}(\bar{y}_s) = \mu$$

$$nV_I(\bar{y}_s) \to \left[1 + E\left(M_i(M_i-1)\tau(M_i)\right)/\mu_M\right]\sigma^2/\mu_M$$

$$nV_{II}(\bar{y}_s) \to \sigma^2/\mu_M$$

where

$$\mu_M = E(M_i) \ , \ \tau(M_i) = E\left((\mu_i-\mu)^2|M_i\right)/\sigma^2$$

Proof :

We use the same notation as in Lemma 5.11

$$E_I(\bar{y}_s) = E_I\left(\sum_{i=1}^{n} M_i\mu(M_i) / \sum_{i=1}^{n} M_i\right)$$

The $\left(M_i\mu(M_i), M_i\right)$ are IID with

$$E\left(M_i\mu(M_i)\right) = \Sigma M\mu(M)h_c(M)$$

$$= \Sigma\mu(M)h_u(M)\Sigma Mh_c(M)$$

$$= \mu\mu_M$$

$$E(M_i) = \Sigma Mh_c(M) = \mu_M$$

Hence $$E_I(\bar{y}_s) \to \mu\mu_M/\mu_M = \mu \quad \text{as} \quad N \to \infty$$

$$E_{II}(\bar{y}_s) = \mu \quad \text{as in Lemma 5.2}$$

$$V_I(\bar{y}_s) = V_I\left(\sum_{i=1}^{n} M_i\mu_i / \sum_{i=1}^{n} M_i\right) + E_I\left(\Sigma M_i\sigma_i^2/(\Sigma M_i)^2\right)$$

In order to evaluate the first term we use a Taylor series expansion of the ratio $g(x,y) = x/y$ where $g_x(x,y) = 1/y$, $g_y(x,y) = -x/y^2$ so

$$nV_I\left(\sum_{i=1}^{n} M_i\mu_i / \sum_{i=1}^{n} M_i\right) \to nV_I(\Sigma M_i\mu_i/n)/\mu_M^2$$

$$+ nV_I(\Sigma M_i/n)\mu^2/\mu_M^2 - 2n\mu\text{cov}(\Sigma M_i\mu_i/n,\ \Sigma M_i/n)/\mu_M^2$$

$$\to V_I(M_i\mu_i)/\mu_M^2 + \sigma_M^2\mu^2/\mu_M^2 - 2\mu\text{cov}(M_i\mu_i, M_i)/\mu_M^2$$

where

$$\sigma_M^2 = \text{Var}(M_i)$$

Hence

$$nV_I(\bar{y}_s) \to V(M_i\mu_i)/\mu_M^2 + \sigma_M^2\mu^2/\mu_M^2 - 2\mu\text{cov}(M_i\mu_i, M_i)/\mu_M^2 + E(M_i\sigma_i^2)/\mu_M^2$$

Now as in (5.25)

$$\sigma^2 = \left[E_I(M_i\mu_i^2) - \mu^2\mu_M + E(M_i\sigma_i^2)\right]/\mu_M$$

Hence

$$n\,V_I(\bar{y}_s) \to \sigma^2/\mu_M + \left[E_I(M_i^2\mu_i^2) - \mu_M^2\mu^2 + \sigma_M^2\,\mu^2 - 2\mu E_I(M_i^2\mu_i) + 2\mu^2\mu_M^2 - E(M_i\mu_i^2) + \mu^2\mu_M\right]/\mu_M^2$$

$$= \sigma^2/\mu_M + \left[E_I(M_i(M_i-1)\mu_i^2) - 2\mu E(M_i(M_i-1)\mu_i) - 2\mu^2\mu_M + \mu_M^2\mu^2 + \sigma_M^2\mu^2 + \mu^2\mu_M\right]/\mu_M^2$$

$$= \sigma^2/\mu_M + E_I\left[M_i(M_i-1)(\mu_i-\mu)^2\right]/\mu_M^2$$

$$= \sigma^2/\mu_M + \sigma^2 E_I\left[M_i(M_i-1)\tau(M_i)\right]\mu_M^2$$

$$= \left[1 + E_I\left[M_i(M_i-1)\tau(M_i)\right]/\mu_M\right]\sigma^2/\mu_M$$

as required.

$$nV_{II}(\bar{y}_s) = nV_{II}(\mu) + nE_{II}\left[\sigma^2 / \sum_{i=1}^{n} M_i\right]$$

$$\rightarrow \sigma^2/\mu_M \qquad \text{as} \quad N \rightarrow \infty$$

Note that Lemma 5.11b also follows by applying a Taylor series linearisation to a special case of Theorem A of Fuller (1975) who obtains the asymptotic moments of the sample mean per psu under Model I.

Lemmas 5.11a and 5.11b suggest a form of robustness of our results to departures from Assumption B when the design is self-weighting.

5.3 Misspecification Effects of Means

In this section we suppose that y_{ij} is univariate (i.e. $p = 1$). The extension to general p is straightforward. We consider

$$T_{Ym} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij}/m_o \qquad (5.26)$$

as an estimator of μ, the mean of f_o defined in (5.3).

Note that T_{Ym} may be expressed in the form T_h of (5.6) with $h(y) = y/m_o$. It follows from Lemma 5.2 that misspecification of Model I as Model II does not introduce any bias under Assumption B. T_{Ym} is unbiased for μ under both models under this assumption.

Lemma 5.12

If Assumption B holds

$$\text{meff}(T_{Ym}|s,\underline{M}) = 1 + \Sigma m_i(m_i-1)\tau_{Ym}(M_i)/m_o \qquad (5.27)$$

where

$$\tau_{Ym}(M_i) = \text{corr}_I(Y_{ij},Y_{ij'}|M_i) \quad j \neq j' \qquad (5.28)$$

Corollary 5.13

If Assumption A holds

$$\text{meff}(T_{Y_m}|s,\underline{M}) = 1 + (m^*-1)\tau_{Ym} \qquad (5.29)$$

Proof:

This follows from (5.15) and (5.17) on the basis of Lemma 5.2.

An alternative representation of $\tau_{Ym}(M)$ may be obtained as in (5.20) by letting

$$\mu_i = E_I(Y_{ij}|\theta_i)$$

$$\sigma_i^2 = V_I(Y_{ij}|\theta_i)$$

$$\sigma_B^2(M_i) = V_I(\mu_i|M_i)$$

$$\sigma_W^2(M_i) = E_I(\sigma_i^2|M_i)$$

Then if B holds

$$\sigma^2 = V_I(Y_{ij}) = \sigma_B^2(M_i) + \sigma_W^2(M_i)$$

and from (5.20) if B holds

$$\tau_{Ym}(M) = \sigma_B^2(M)/\sigma^2 \qquad (5.30)$$

$$= 1 - \sigma_W^2(M)/\sigma^2 \qquad (5.31)$$

If A holds $\sigma_B^2(M) = \sigma_B^2$, $\sigma_W^2(M) = \sigma_W^2$ and

$$\tau_{Ym} = \sigma_B^2/\sigma^2$$

The above results depend fundamentally on Assumption B. Recall that the major component of Assumption B is that the marginal distribution of Y_{ij} does not depend on M_i (except that $j \leq M_i$). In particular for there to be no misspecification bias in T_{Ym} we need $E(Y_{ij}|M_i)$ free of M_i . This might be checked diagnostically by plotting

$$\bar{y}_i = \sum_{j=1}^{m_i} y_{ij}/m_i$$

against M_i . If B holds the regression function $E(\bar{y}_i|M_i)$ should not depend on M_i.

We give in Figures 5.1 - 5.3 such plots for a National Survey of Attainment conducted in Wales in 1960. The sample design was a stratified cluster sampling design where clusters were schools. The strata were defined by geographical region , type of school and sex of school. The schools were selected from a Ministry of Education list by using a fixed sampling interval for each stratum. For sampled schools all children aged over 14 were selected, i.e. $m_i = M_i$. After excluding children with missing values on the variables of interest a total of 3053 children remained, divided into $n = 50$ clusters. The M_i's differed greatly ranging from 5 to 136 ($\bar{m} = 61.1$, $m^* = 77.8$).

The variables considered were

AM : Attitude to Mathematics - 7 point scale.

TM : Mathematics Test - 85 item test, scores out of 85

W : Welsh reading test - 35 item test, scores out of 35.

The effect of stratification was investigated but did not affect the conclusions and strata are not indicated below.

Figure 5.1 : National Survey - AM-Attitude to Mathematics

Figure 5.2 : National Survey
TM - Mathematics Test

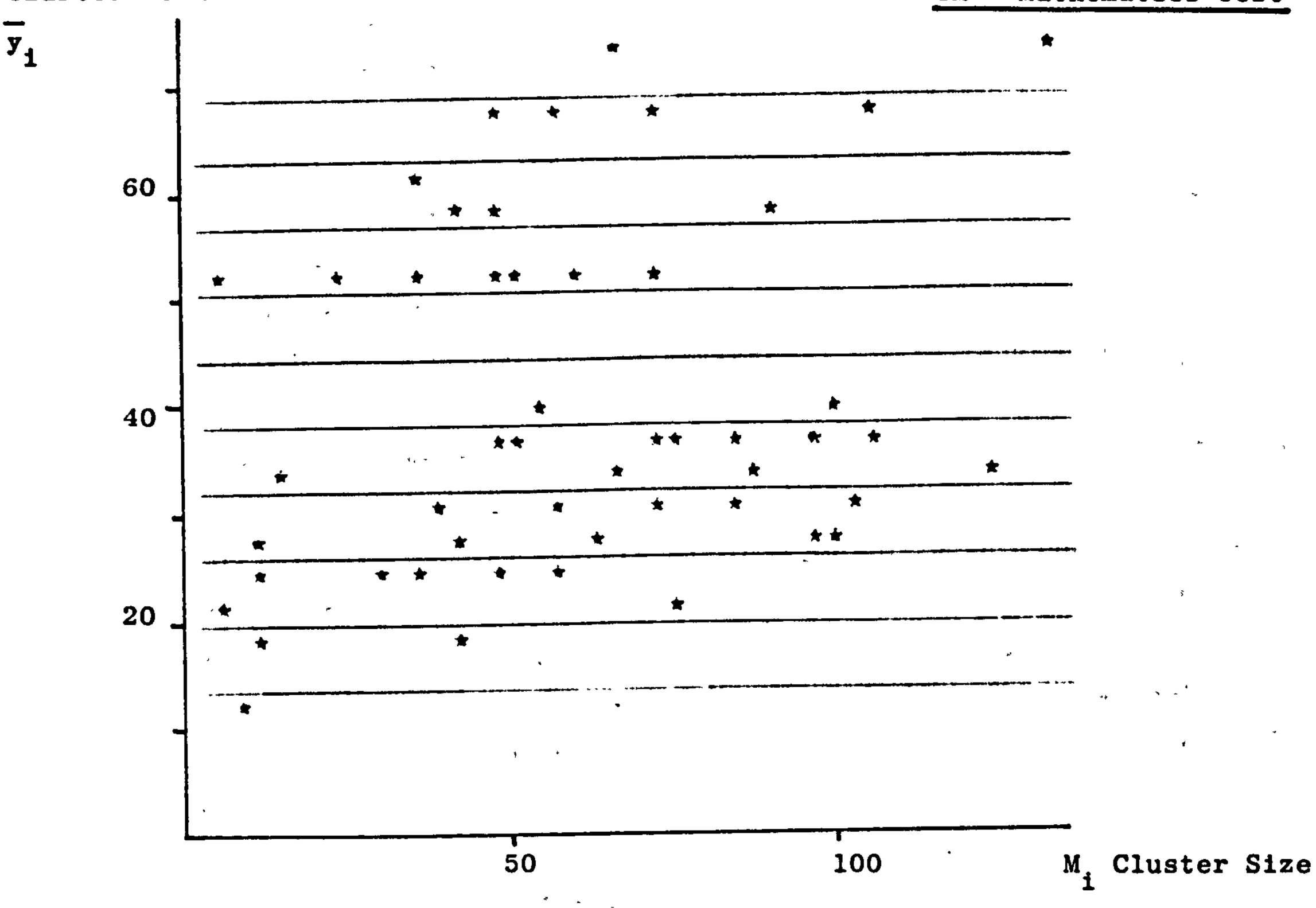
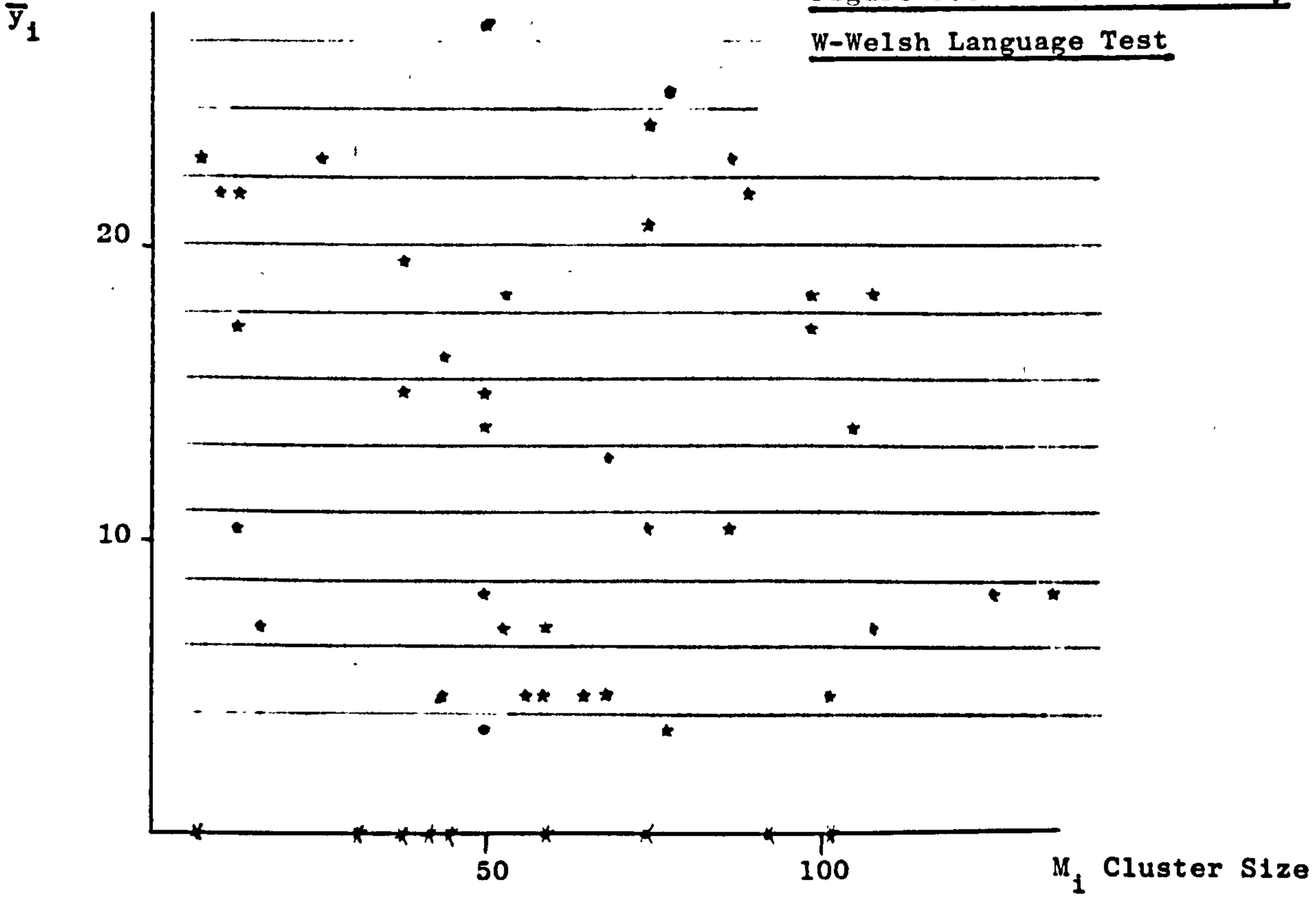


Figure 5.3 - National Survey
W-Welsh Language Test



Firstly, we note that the (conditional) standard errors of the $\bar{y}_i$ as estimators of the μ_i are relatively small being approximately $1.2\sqrt{M_i}$, $15.2\sqrt{M_i}$ and $6.3\sqrt{M_i}$ for figures 5.1 to 5.3 respectively. Indeed the standard errors are less than the minimum calibration of the computer plotting routine if $M_i \geq 36$, 24 or 29 for Figure 5.1, 5.2 or 5.3 respectively. There is no evidence in Figure 5.1 of $E(\bar{y}_i|M_i)$ depending on M_i. Note that the dependence of $V(\bar{y}_i|M_i)$ on M_i is permissible under B. In Figure 5.2 the regression of $\bar{y}_i$ on M_i appears to increase slightly with M_i. The clustering of schools into selective grammar schools with high $\bar{y}_i$ and other schools with low $\bar{y}_i$ is apparent. Assumption B appears (just) untenable here because mathematical ability is related to school size (probably because of urban/rural differences). Figure 5.3 again does not appear to present any evidence against Assumption B. Note that in seven schools/clusters all children scored zero on the Welsh test (and in two schools almost all scored zero), indicating the strength of the inter-cluster differences on this variable.

It is also interesting to be able to check the validity of Assumption A since the form of the misspecification effect in Corollary 5.13 is rather simpler than that in Lemma 5.12. It is clear from (5.31) that a sufficient condition for the simpler form of misspecification effect to hold is that $\sigma_W^2(M)$ does not depend on M. A natural diagnostic check of this condition is obtained by plotting

$$\hat{\sigma}_i^2 = \sum_{j=1}^{m_i} (y_{ij}-\bar{y}_i)^2/(m_i-1) \qquad (5.32)$$

against M_i. If A holds then the regression function $E_I(\hat{\sigma}_i^2|M_i) = E_I(\sigma_i^2|M_i) = \sigma_W^2(M_i)$ should not depend on M_i. Such plots for the National Survey of Attainment data are given in Figures 5.4-5.6.

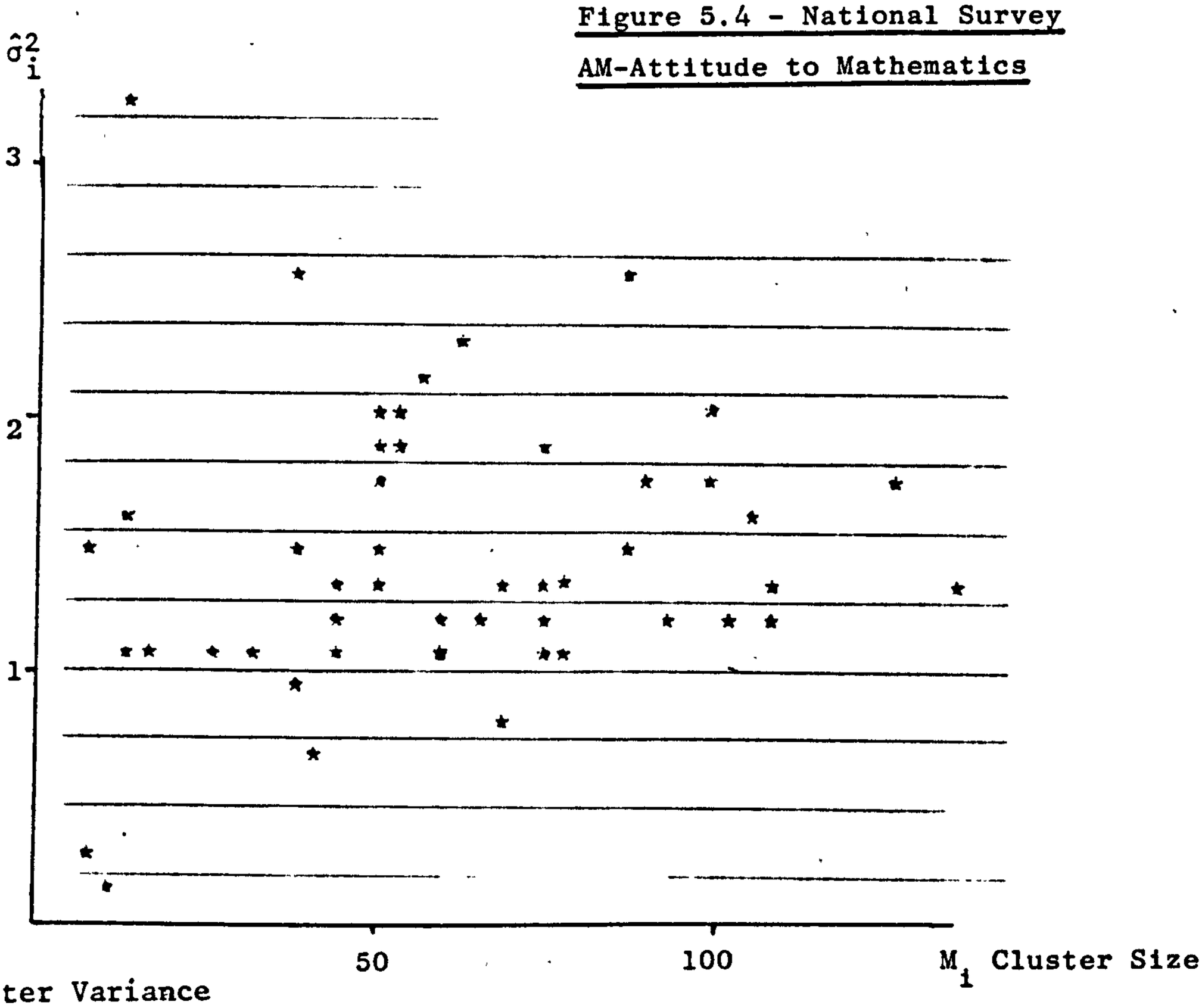


Figure 5.4 - National Survey

AM-Attitude to Mathematics

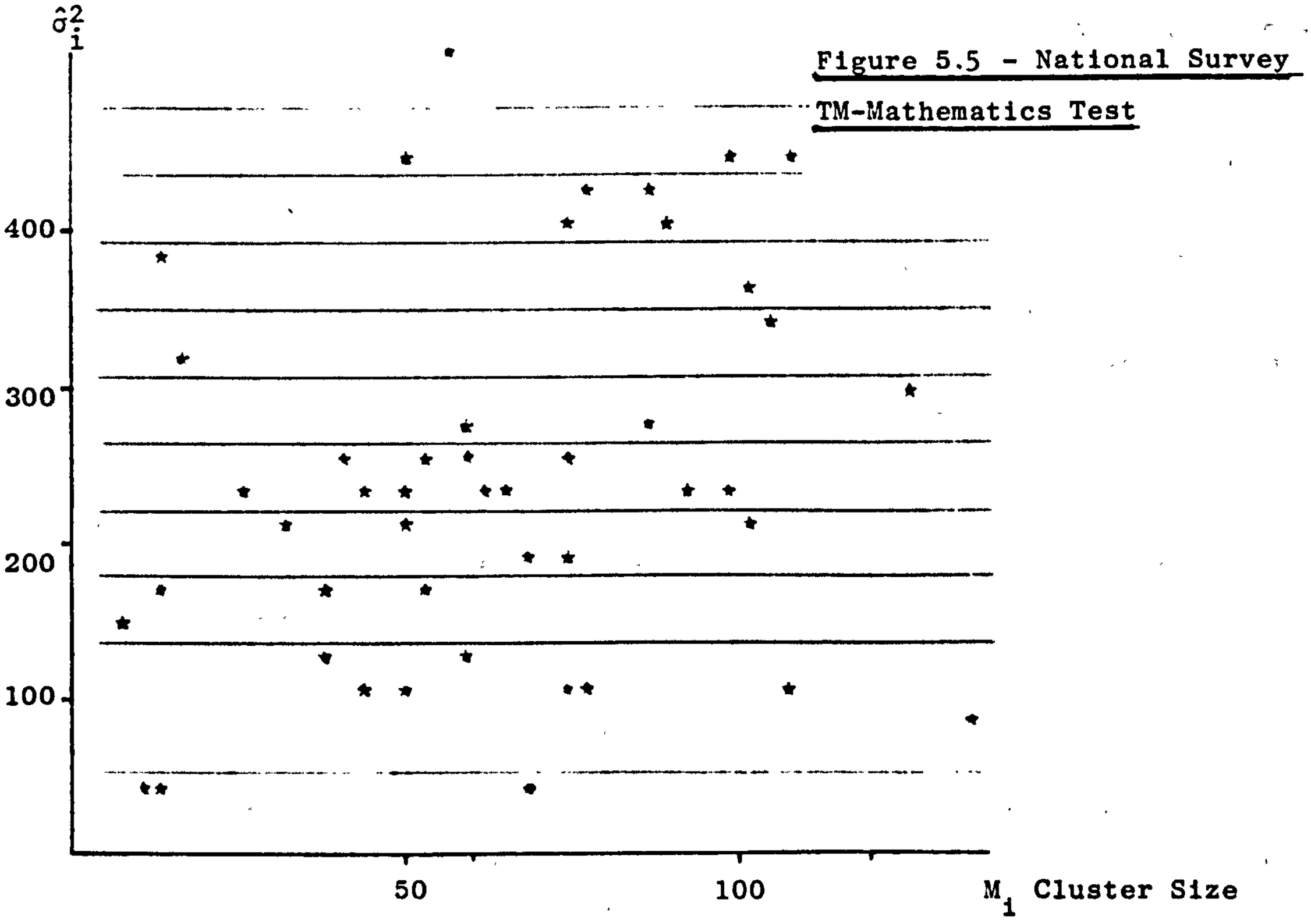


Figure 5.5 - National Survey

TM-Mathematics Test

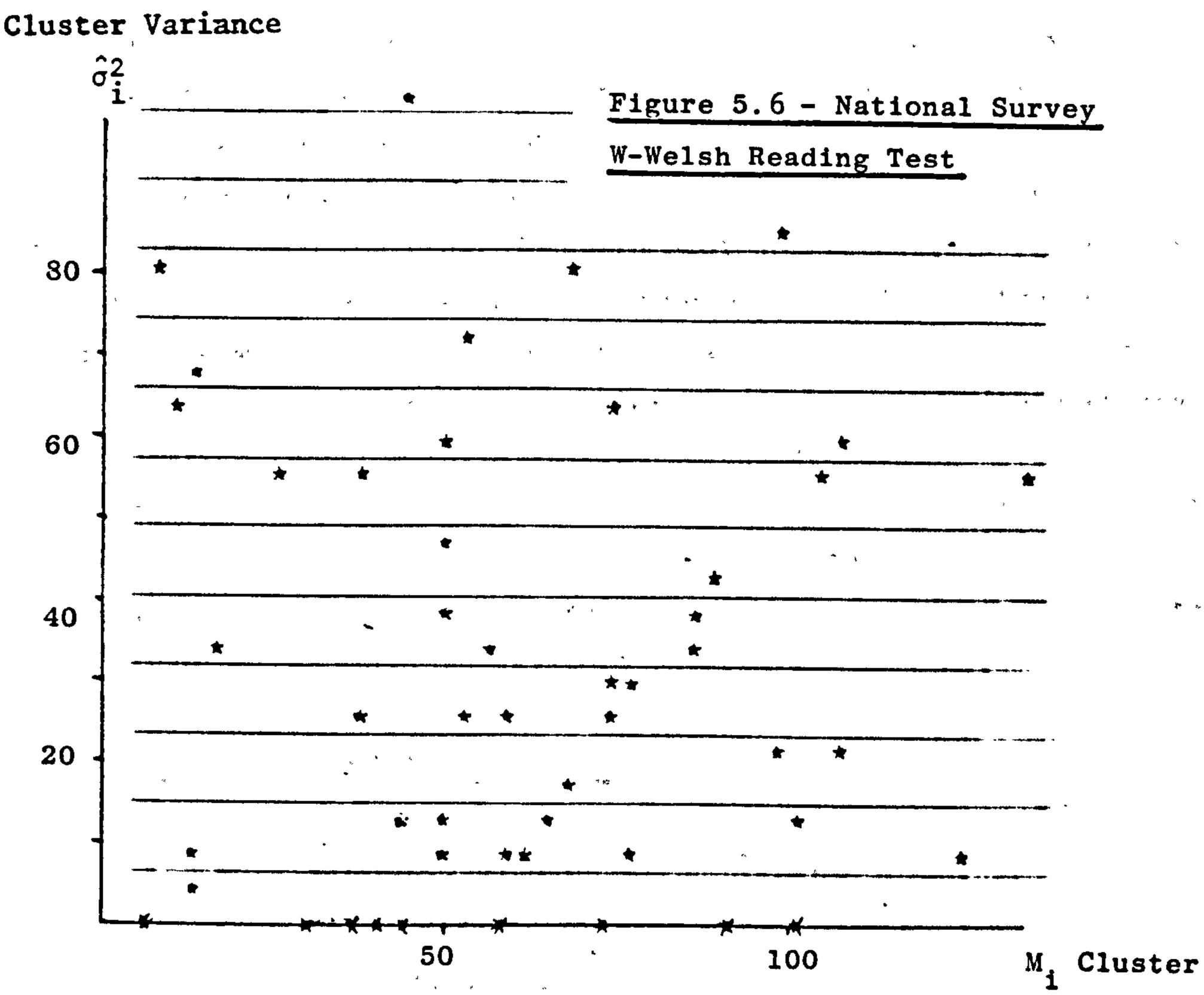


Figure 5.6 - National Survey W-Welsh Reading Test

In none of these plots does the regression function clearly deviate from the horizontal. This is in contrast to what we would expect if the intra-cluster correlation were in fact a decreasing function of M_i (see Section 5.1) in which case we would expect $E(\sigma_i^2|M_i)$ to increase monotonically to an asymptote, σ^2

In the example above, Assumption A seemed reasonable. In general, however, $\tau_{Ym}(M)$ will depend on M, say

$$\tau_{Ym}(M) = f(M) \qquad (5.33)$$

In Section 5.1 we distinguished between the *intra-survey* dependence of $\tau_{Ym}(M)$ and the *inter-survey* dependence. We noted that there has been a certain amount of empirical investigation of the latter dependence but not, to our knowledge, of the former. We noted also that these two dependences might be very

different essentially because in the former case the M_i might be 'proxies' for background variables highly related to the y_{ij}. We also argued, however, that stratification might account for such relations between the M_i and y_{ij}. For example, in the 1975 Family Expenditure Survey the average size of psu's (M_i) per stratum ranged from 3,800 in rurual Southern Scotland to 241,000 in Greater London. Across the whole population we might find that the intracluster correlation was higher on some variables in a London psu than in a Southern Scottish psu of much smaller size. However, within strata we might find that the intra-stratum dependence of τ_{Ym} on M was similar to the inter-survey dependence. This would be the case if the cluster sizes were 'randomly distributed' irrespective of the y_{ij} values. On the basis of this argument we now investigate some of the literature on the inter-survey form of $f(M)$.

(i) $\underline{f_1(M) = a + b/M \quad ; \; a, \; b > 0}$

This is suggested by Cochran (1963, p.256).

(ii) $\underline{f_2(M) = aM^{-b} \quad ; \; a, \; b > 0}$

This is suggested by Hansen et al (1953). It originates from work by Smith (1938) who performed an empirical investigation of crop experiments where Y referred to yield and the clusters were plots of size M_i. Further empirical evidence from crop experiments was provided by Mahalanobis (1944).

Let

$$\bar{Y}_i = \sum_{j=1}^{M_i} Y_{ij}/M_i$$

Let

$$g(M_i) = V_I(\bar{Y}_i|M_i) \tag{5.34}$$

$$= \sigma_W^2(M_i)/\, M_i + \sigma_B^2(M_i)$$

$$= (1 + (M_i-1)\tau_{Ym}(M_i))\sigma^2/M_i \tag{5.35}$$

On the basis of his empirical evidence Smith (1938) proposed the following 'law':

$$g(M_i) = \sigma^2 M_i^{-b} , \quad 0 \leq b \leq 1$$

This model and its generalisation

$$g(M_i) = a\sigma^2 M_i^{-b} \tag{5.36}$$

have been widely used in the survey sampling literature (e.g. Cochran, 1977, p.256). (Note that the special value $a = 1$ follows from $g(1) = \sigma^2$). Combining (5.35) and (5.36) we have

$$\tau_{Ym}(M) = (aM^{-b+1}-1)/(M-1)$$

$$\rightarrow aM^{-b} = f_2(M) \quad \text{as} \quad M \rightarrow \infty$$

Such a connection between Smith's (1938) law and Hansen et al's (1953) model was demonstrated by Brewer et al (1977). Note also that if $\tau_{Ym}(M) = f_1(M)$ is substituted into (5.35) then

$$g(M_i) = \left(1 + (M_i-1)(a+b/M_i)\right)\sigma^2/M_i$$

$$= \left(a + (1-a+b)/M_i - b/M_i^2\right)\sigma^2$$

which Cochran (1963, p.256) argues may also be approximated by

$$g(M_i) = cM_i^{-d} \quad \text{where} \quad 0 \leq d \leq 2$$

as in (5.36)

(iii) $f_3(M) = 1 - aM^b \quad ; 0 < a, b < 1$

On the basis of farm survey data, Jessen (1942) proposed a model, which in our notation may be written

$$\sigma_W^2(M) = cM^b$$

This implies

$$f(M) = (\sigma^2 - cM^b)/\sigma^2$$

$$= f_3(M) \qquad \text{where} \quad a = c/\sigma^2$$

$f_3(M)$ gave a better fit to Jessen's data than $f_2(M)$ but f_3 has the theoretical disadvantage, noted by Hendricks (1944), that $f_3(M) \to -\infty$ as $M \to \infty$, violating the constraint $|\tau| \leq 1$.

As an illustration, Table 5.1 contains some estimated values of $\tau_{Ym}(M_i)$ and corresponding M_i from Hansen et al (1953) for variables with low, medium and high intracluster correlation.

Table 5.1 Intra-cluster correlations of selected characteristics of selected cities over 100,000

M_i	$\tau_{Ym}(M_i)$		
	Males 25-34	Males in Labour Force	Average Rental Value
3	.045	.12	.45
9	.026	.10	.36
27	.018	.07	.25
62	.0079	.03	.12

These values are plotted in Figures 5.7-5.9. For this data it appears that the best fit is obtained in Figure 5.9 for $f_3(M)$, the least theoretically attractive function.

Figure 5.7

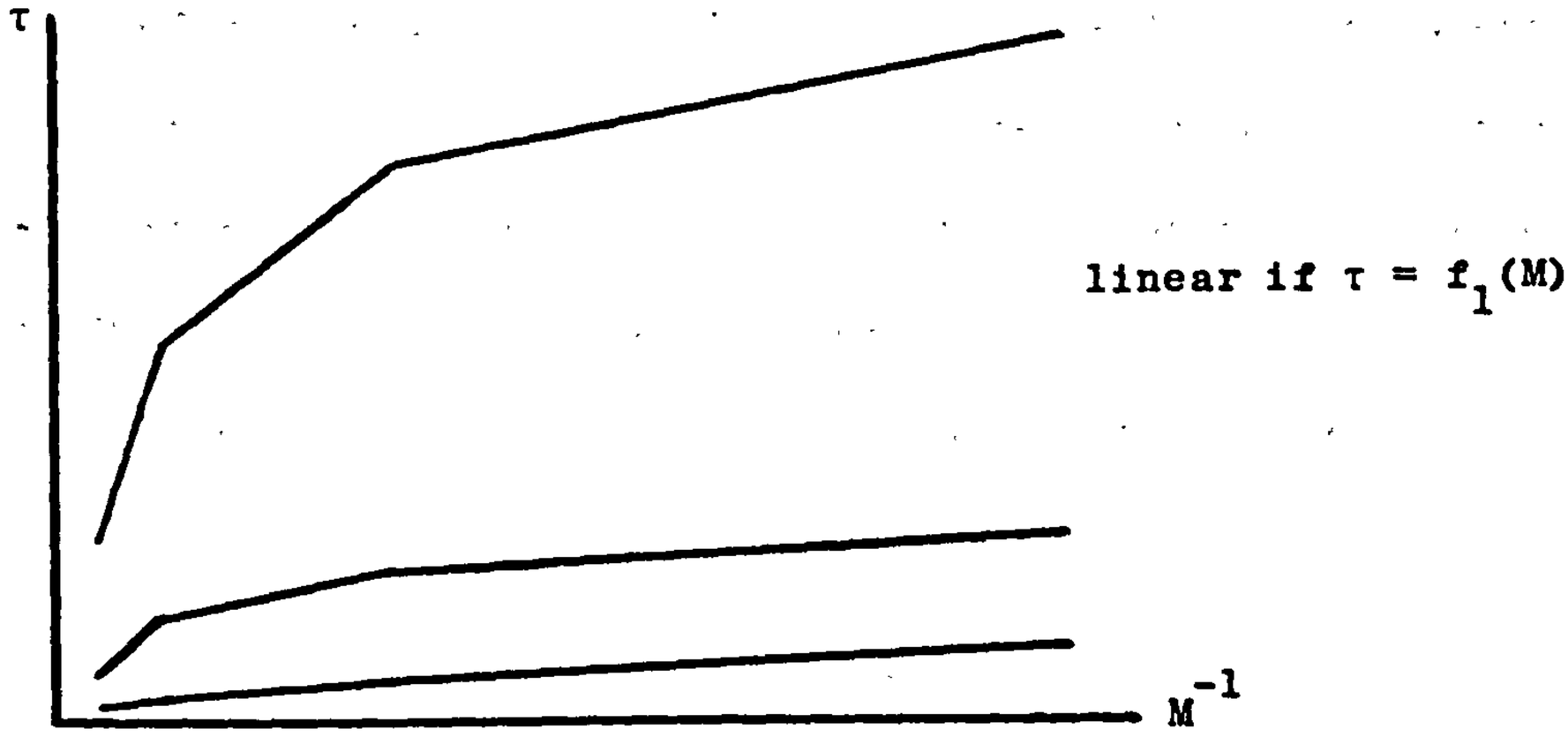


Figure 5.8

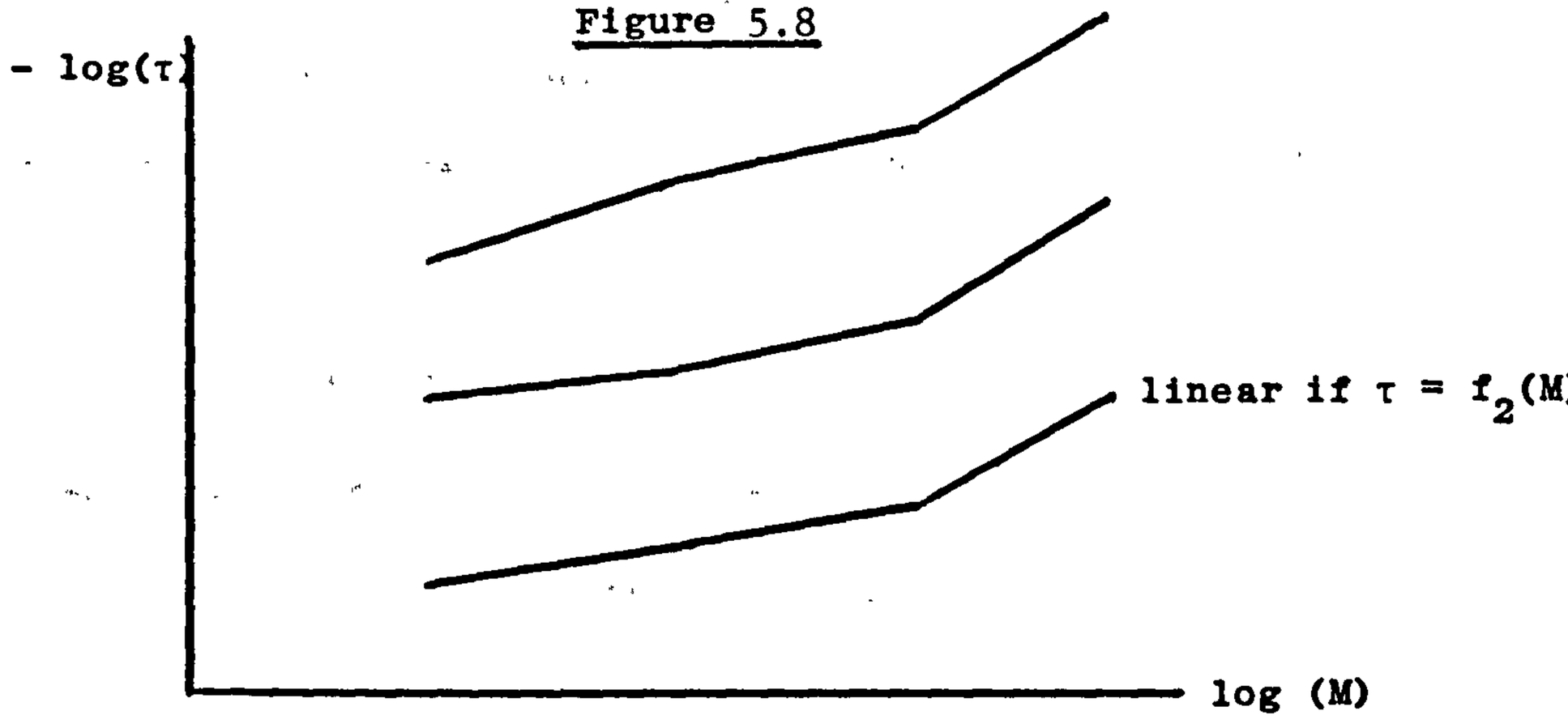
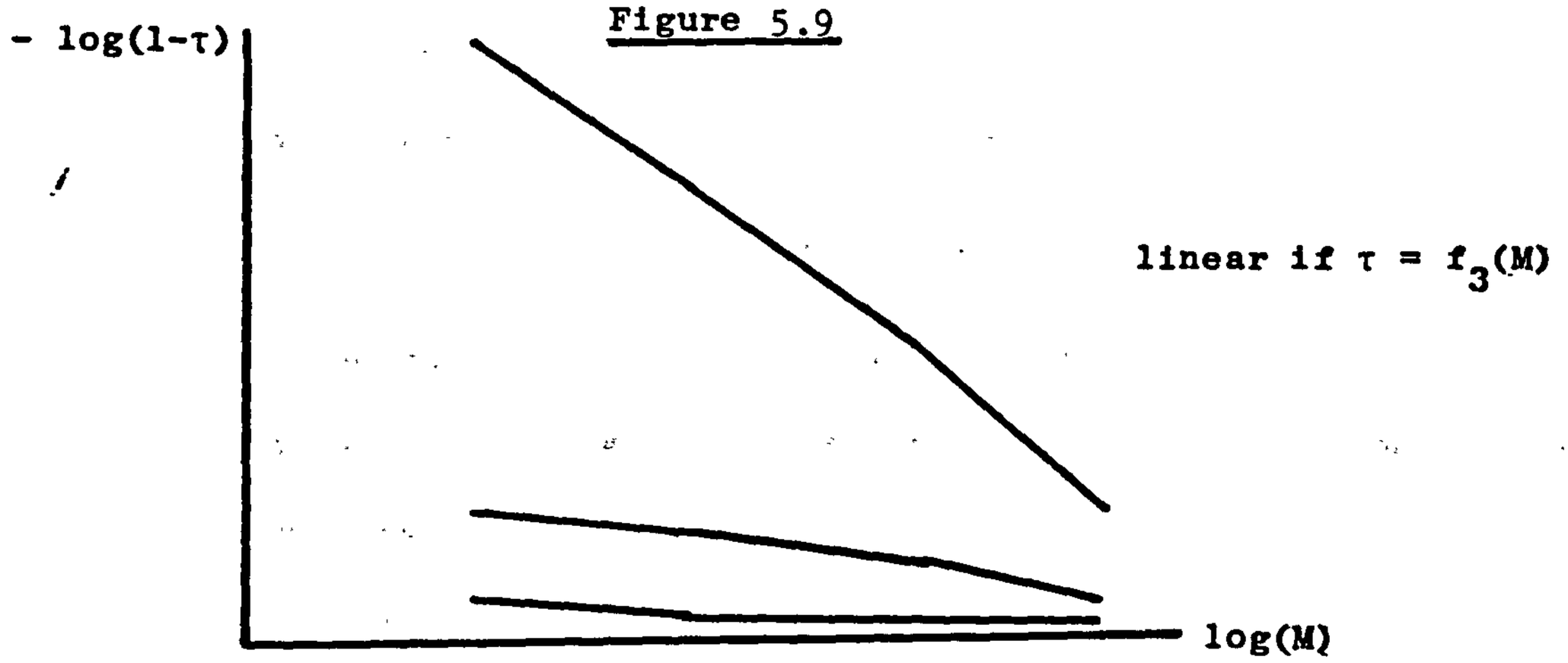


Figure 5.9



In the case of area sampling it is natural to view Model I as a spatial process (Cochran, 1946; Whittle 1956, 1962) where the intra-cluster correlation is related to the spatial correlation between members of the same cluster. Let $Y(\underline{x})$ be a random variable indexed by the planar coordinates $\underline{x}$. We assume the process is stationary:

$$E(Y(\underline{x})) = \mu \qquad \text{for all } \underline{x}$$

$$V(Y(\underline{x})) = \sigma^2 \qquad \text{for all } \underline{x}$$

and isotropic:

$$\text{corr}(Y(\underline{x}), Y(\underline{x}')) = \rho(s)$$

for all $\underline{x}$ and $\underline{x}'$ a distance s apart.

Consider a cluster defined by a geographical region Ω_i of area A_i.

$$A_i = \int_{\Omega_i} d\underline{x}$$

Let $\mu_i = \int_{\Omega_i} Y(\underline{x})\, d\underline{x}/A_i$,

the mean value of Y in the cluster.

Then under B, the intracluster correlation in Ω_i is

$$\tau_i = V(\mu_i)/\sigma^2$$
$$= \int_{\Omega_i}\int_{\Omega_i} \text{cov}(Y(\underline{x}), Y(\underline{x}')\, d\underline{x}\, d\underline{x}'/\sigma^2 A_i^2 \qquad (5.37)$$

Let α_i be the maximum distance between two points in Ω_i and let $K_i(s)$ denote the distribution of distances between two points chosen randomly in Ω_i. Then, changing variables, we have from (5.37)

$$\tau_i = \int_o^{\alpha_i} \rho(s)\, K_i(s)\, ds \qquad (5.38)$$

This formula indicates how τ_i is a weighted mean of the spatial correlations between points in Ω_i. Whittle (1956) suggests considering a class of regions Ω_i of the same shape so that the size of each region is specified by α_i.

Then $A_i = A(\alpha_i) = \alpha_i^2 A(1)$

$K_i(s) = K(s/\alpha_i)/\alpha_i$

and, changing variables in (5.38) to $t = s/\alpha_i$, we have

$$\tau_i = \int_o^1 \rho(\alpha_i t)\, K(t)\, dt \qquad (5.39)$$

Hence, for example, if

$$\rho(s) = as^{-b}$$

then $$\tau_i = a\alpha_i^{-b} \int_o^1 t^{-b} K(t)\, dt$$

$$\propto \alpha_i^{-b}$$

or if $\rho(s) = O(s^{-b})$

then $\tau_i = O(\alpha_i^{-b})$

In general τ_i is a convolution of ρ and K. If ρ has the more familiar exponential form

$$\rho(s) = e^{-\lambda s}$$

then $$\tau_i = \int_o^1 e^{-\lambda\alpha_i t} K(t)\, dt ,$$

a Laplace transform of K. If, for example, K is a Gamma distribution with parameters β and r then

$$\tau_i \propto (\lambda\alpha_i + \beta)^{-r}$$

Alternatively suppose Ω_i is the one-dimensional strip $(0,\alpha_i)$ then

$$\tau_i = 2\int_o^{\alpha_i} \int_o^t e^{-\lambda(t-s)}\, ds\, dt/\alpha_i^2$$

$$= 2\int_o^{\alpha_i} (1 - e^{-\lambda t})\, dt/\alpha_i^2\lambda$$

$$= 2\,\alpha_i^{-1}/\lambda - 2(1 - e^{-\lambda\alpha_i})\,\alpha_i^{-2}/\lambda$$

$$\to 2\alpha_i^{-1}/\lambda \quad \text{as} \quad \alpha_i \to \infty$$

Hence the functional form of τ may be quite different to that of ρ.

If ρ and the population density are uniform across clusters (within strata). Then

$$\alpha_i \propto M_i^{\frac{1}{2}}$$

and for example if

$$\rho(s) = as^{-b}$$

then

$$\tau_i \propto M_i^{-b/2}$$

as in $f_2(M)$.

The population density $\delta_i = M_i/\alpha_i^2$ may, however, be related to M_i e.g. in an area with high δ_i an interviewer may be given a random sample of addresses from a cluster with high M_i since the cluster area will be relatively small and vice versa. The intra-cluster correlation will then only have the above form if $\rho_i(s) = \rho(s\sqrt{\delta_i})$ since in this case

$$\tau_i = \int_0^1 \rho_i(\alpha_i t)\ K(t)dt$$

$$= \int_0^1 \rho_i(\sqrt{M_i/\delta_i}\ t)\ K(t)dt$$

$$= \int_0^1 \rho(\sqrt{M_i}t)\ K(t)\ dt$$

In general the functional form $f(M)$ may be investigated by plotting $\hat{\sigma}_i^2$ (defined in 5.32) against M_i as in Figures 5.4-5.6.

$$E_I(\hat{\sigma}_i^2 | M_i) = \sigma_W^2(M_i)$$

$$= \sigma^2(1 - f(M_i))$$

The estimation of σ^2 is discussed in Section 5.4 and Chapter 6. Note that for weighted least squares fitting

$$V_I(\hat{\sigma}_i^2|M_i) = V_I(\sigma_i^2|M_i) + E_I(V(\hat{\sigma}_i^2|\theta_i)|M_i)$$

$$= V_I(\sigma_i^2|M_i) + O(m_i^{-1})$$

Hence the weights depend mainly on the model assumptions about $V_I(\sigma_i^2|M_i)$.

5.4 Misspecification Effects of Variances

As in Section 5.3, we suppose y_{ij} is univariate.

Let

$$T_{Yv} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij}-T_{Ym})^2/(m_o-1) \qquad (5.40)$$

where T_{Ym} is defined in (5.26)

We consider T_{Yv} as an estimator (the standard estimator) of σ^2, the variance of f_o defined in (5.3). We could express T_{Yv} as a non-linear function , $g(\underline{T})$, of additive statistics of the form T_h in (5.6) as say in Krewski and Rao (1981). But we prefer instead to approximate T_{Yv} by

$$T_{Yv}^{\sim} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} h_v(y_{ij})$$

where

$$h_v(y) = (y - \mu)^2/(m_o-1)$$

and μ is the mean of f_o .

We then approximate the misspecification effect of T_{Yv} by that of $T_{Yv}^{\sim}$. Note that the fact that $T_{Yv}^{\sim}$ is not an observable statistic does not matter since we are only interested in the theoretical properties of T_{Yv} . We suspect that the moments of $T_{Yv}^{\sim}$ are a better approximation to the moments of T_{Yv} than the usual Taylor series approximation but we do not intend to prove this. We shall however demonstrate the asymptotic equivalence, in a certain sense, of the misspecification effects of T_{Yv} by $T_{Yv}^{\sim}$ in Theorem 5.23. The main advantage of $T_{Yv}^{\sim}$ is that it is of the additive form (5.6) and hence we may use the simple results of Section 5.2.

From Lemma 5.2 it follows that misspecification of Model I as Model II does not introduce any bias into $T_{Yv}^{\sim}$ under Assumption B. In fact

$$E_I(T_{Yv}^{\sim}|s,\underline{M}) = E_{II}(T_{Yv}^{\sim}|s,\underline{M}) = m_o\sigma^2/(m_o-1) \qquad (5.40a)$$

Lemma 5.14

If Assumption B holds

$$\text{meff}(T_{Yv}^{\sim}|s,\underline{M}) = 1 + \sum_{i=1}^{n} m_i(m_i-1)\tau_{Yv}^{\sim}(M_i)/m_o \qquad (5.41)$$

where

$$\tau_{Yv}^{\sim}(M_i) = \text{corr}_I\left[(Y_{ij}-\mu)^2, (Y_{ij'}-\mu)^2|M_i\right] , \quad j \neq j' \qquad (5.42)$$

Corollary 5.15

If Assumption A holds

$$\text{meff}(T_{Yv}^{\sim}|s,\underline{M}) = 1 + (m^*-1)\tau_{Yv}^{\sim} \qquad (5.43)$$

Proof:

These results follow from (5.15) and (5.17)

In order to express $\tau_{Yv}^{\sim}(M_i)$ in terms of θ_i we introduce some notation.

Let

$$k_{3i} = E_I\left[(Y_{ij}-\mu_i)^3|\theta_i\right]$$

$$k_{4i} = E_I\left[(Y_{ij}-\mu_i)^4|\theta_i\right] - 3\sigma_i^4$$

$$\gamma(M_i) = V_I(\sigma_i^2|M_i)$$

$$k_{4W}(M_i) = E_I(k_{4i}|M_i)$$

$$k_{4B}(M_i) = E_I\left[(\mu_i-\mu)^4|M_i\right] - 3\sigma_B^4(M_i)$$

$$c_1(M_i) = \text{cov}_I\left[(\mu_i-\mu)^2, \sigma_i^2|M_i\right]$$

$$c_2(M_i) = \text{cov}_I\left[\mu_i,k_{3i}|M_i\right]$$

Lemma 5.16

If B holds the fourth cumulant of f_o , k_4 , obeys the following identity for any M_i.

$$k_4 = k_{4W}(M_i) + k_{4B}(M_i) + 3\gamma(M_i) + 6c_1(M_i) + 4c_2(M_i)$$

Proof:

$$k_4 = E_I\left[(Y_{ij}-\mu)^4 | M_i\right] - 3\sigma^4$$

$$= E_I\Big[E_I\big((Y_{ij}-\mu_i)^4 + 4(Y_{ij}-\mu_i)^3(\mu_i-\mu) + 6(Y_{ij}-\mu_i)^2(\mu_i-\mu)^2$$

$$+ 4(Y_{ij}-\mu_i)(\mu_i-\mu)^3 + (\mu_i-\mu)^4 | \theta_i\big) | M_i\Big]$$

$$- 3\sigma^4$$

$$= E_I\left[k_{4i} + 3\sigma_i^4 + 4k_{3i}(\mu_i-\mu) + 6\sigma_i^2(\mu_i-\mu)^2 + (\mu_i-\mu)^4 | M_i\right] - 3\sigma^4$$

$$= k_{4W}(M_i) + 3\gamma(M_i) + 3\sigma_W^4(M_i) + 4c_2(M_i) + 6\sigma_W^2(M_i)\sigma_B^2(M_i)$$

$$+ 6c_1(M_i) + k_{4B}(M_i) + 3\sigma_B^4(M_i) - 3\left(\sigma_B^2(M_i) + \sigma_W^2(M_i)\right)^2$$

$$= k_{4W}(M_i) + k_{4B}(M_i) + 3\gamma(M_i) + 6c_1(M_i) + 4c_2(M_i)$$

Lemma 5.17

If B holds

$$\tau_{Y\tilde{v}}(M_i) = \left[2\sigma_B^4(M_i) + k_{4B}(M_i) + 2c_1(M_i) + \gamma(M_i)\right]/(2\sigma^4+k_4)$$

Proof :

$$\tau_{Y\tilde{v}}(M_i) = \text{corr}_I\left[(Y_{ij}-\mu)^2,\ (Y_{ij'}-\mu)^2 \middle| M_i\right]$$

$$= \frac{\text{var}_I\left[E\left[(Y_{ij}-\mu)^2|\theta_i\right]|M_i\right]}{\text{var}_I\left[(Y_{ij}-\mu)^2|M_i\right]} \qquad (5.44)$$

$$= \text{var}_I\left[\sigma_i^2 + (\mu_i-\mu)^2|M_i\right]/(2\sigma^4 + k_4)$$

$$= \left[2\sigma_B^4(M_i) + k_{4B}(M_i) + 2c_1(M_i) + \gamma(M_i)\right]/(2\sigma^4 + k_4)$$

Corollary 5.18

If A holds

$$\tau_{Y\tilde{v}} = (2\sigma_B^4 + k_{4B} + 2c_1 + \gamma)/(2\sigma^4 + k_4) \qquad (5.45)$$

We shall consider these results later in this Section but initially we consider diagnostic checks of Assumptions B and A. In order for there to be no misspecification bias we need $E(h_v(Y_{ij})|M_i)$ to be free of M_i. This requirement may be checked by plotting

$$(m_o-1)\hat{\bar{h}}_{vi} = (m_o-1)\sum_{j=1}^{m_i}\hat{h}_v(y_{ij})/m_o \qquad (5.46)$$

$$= \sum_{j=1}^{m_i}(y_{ij}-\hat{\mu})^2/m_o$$

where

$$\hat{\mu} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}/m_o \qquad (5.47)$$

against M_i. If B holds the regression function $E((m_o-1)\hat{\bar{h}}_{vi}|M_i)$ should not depend on M_i (assuming the effect of estimating μ is negligible). Such plots are given in Figures 5.10-5.12 for the National Survey of Attainment data. .

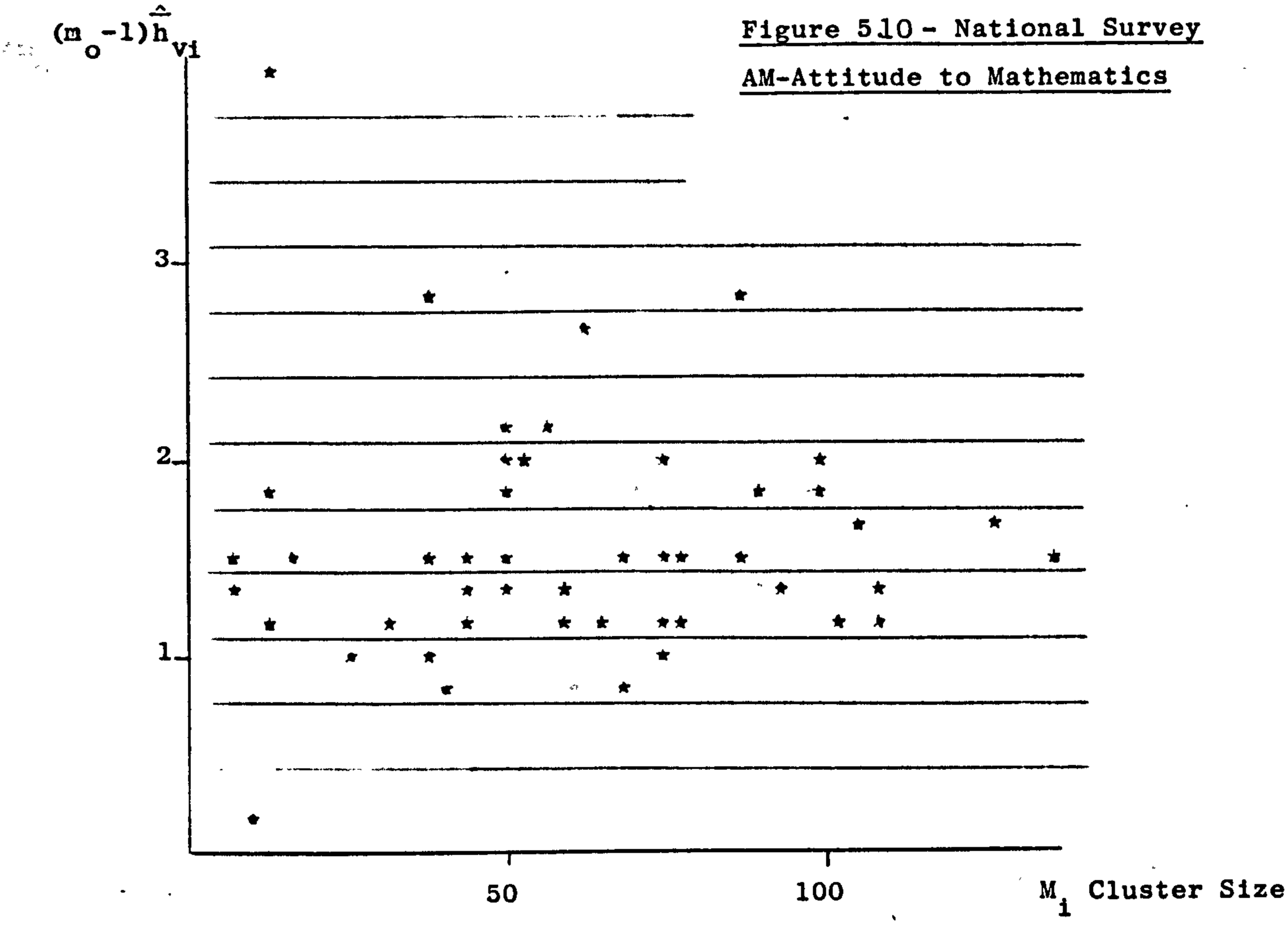


Figure 5.10 - National Survey

AM-Attitude to Mathematics

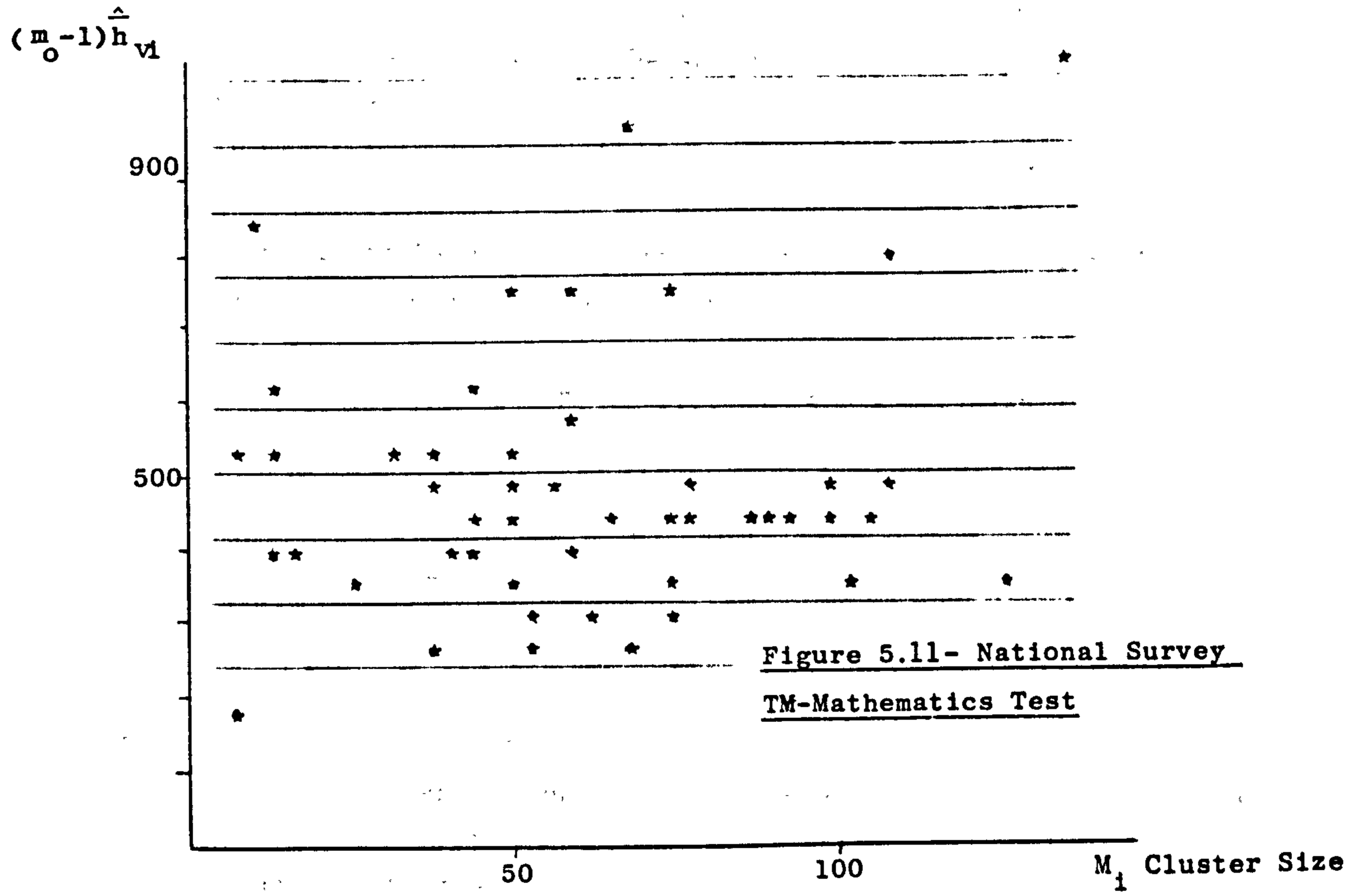


Figure 5.11- National Survey

TM-Mathematics Test

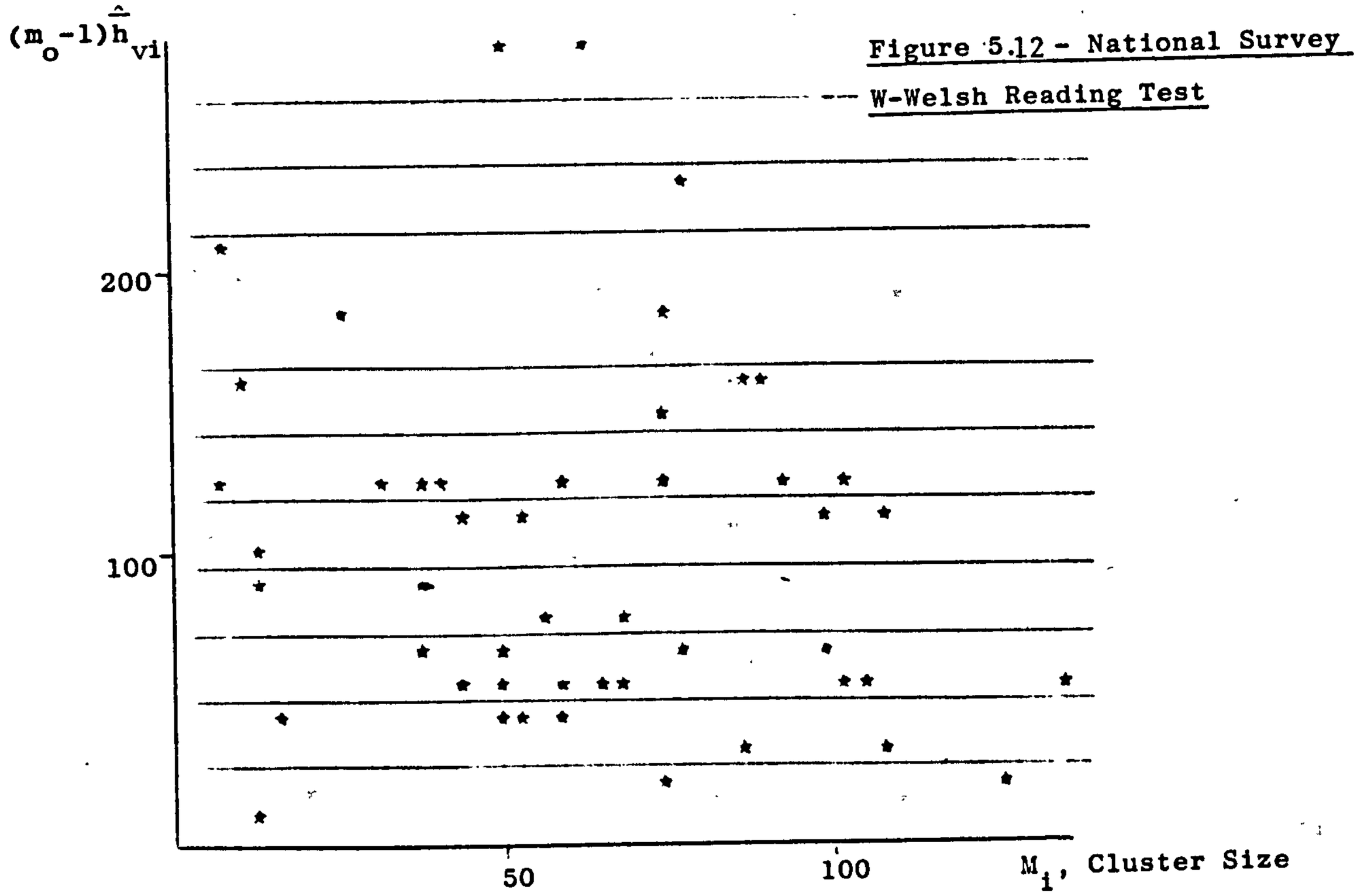


Figure 5.12 - National Survey W-Welsh Reading Test

As in Figure 5.1, there is no obvious evidence in Figure 5.10 that Assumption B is violated. The trend in Figure 5.2 is no longer evident in Figure 5.11 nor is the selective/non-selective school clustering. Figure 5.12 seems broadly similar to Figure 5.3 although there is possibly a decrease in the regression function.

Again the form of the misspecification effect in Corollary 5.15 is rather simpler when Assumption A holds. Note that from (5.44)

$$\tau_{Y\tilde{v}}(M_i) = 1 - E_I\left[V_I\left[(Y_{ij}-\mu)^2|\theta_i\right]|M_i\right]/(2\sigma^4+k_4) \qquad (5.48)$$

An estimator (predictor) of $V_I\left[(Y_{ij}-\mu)^2|\theta_i\right]$ is

$$\hat{\sigma}^2_{vi} = \sum_{j=1}^{m_i}\left[(y_{ij}-\hat{\mu})^2 - \sum_{k=1}^{m_i}(y_{ik}-\hat{\mu})^2/m_i\right]^2/(m_i-1) \qquad (5.49)$$

Hence the validity of Corollary 5.15 may be checked by plotting $\hat{\sigma}^2_{vi}$ against M_i. If Assumption A holds the regression function $E(\hat{\sigma}^2_{vi}|M_i)$ should not depend on M_i (assuming the effect of estimating μ is negligible). Such plots for the National Survey of Attainment data are given in Figures 5.13 - 5.15.

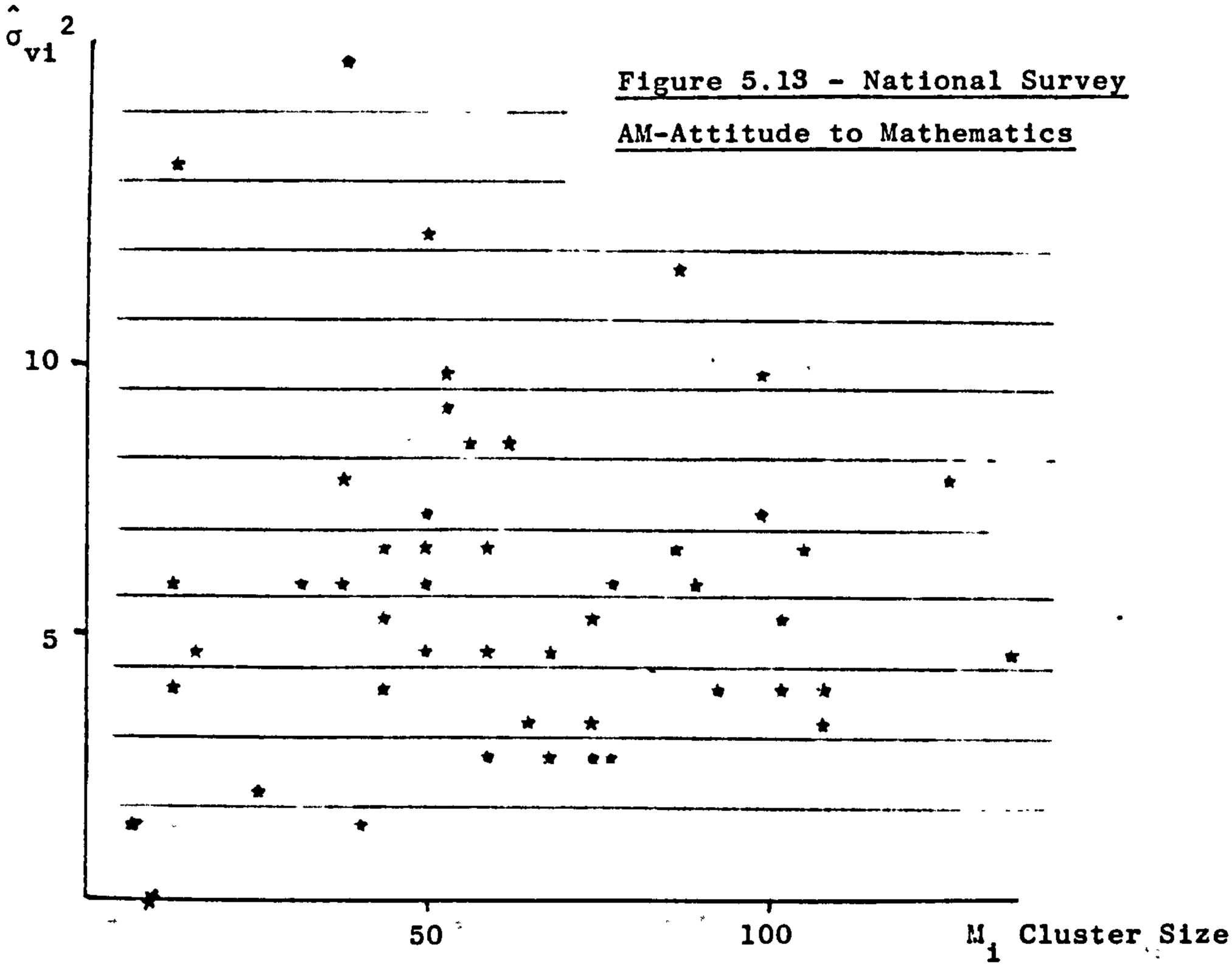


Figure 5.13 - National Survey
AM-Attitude to Mathematics

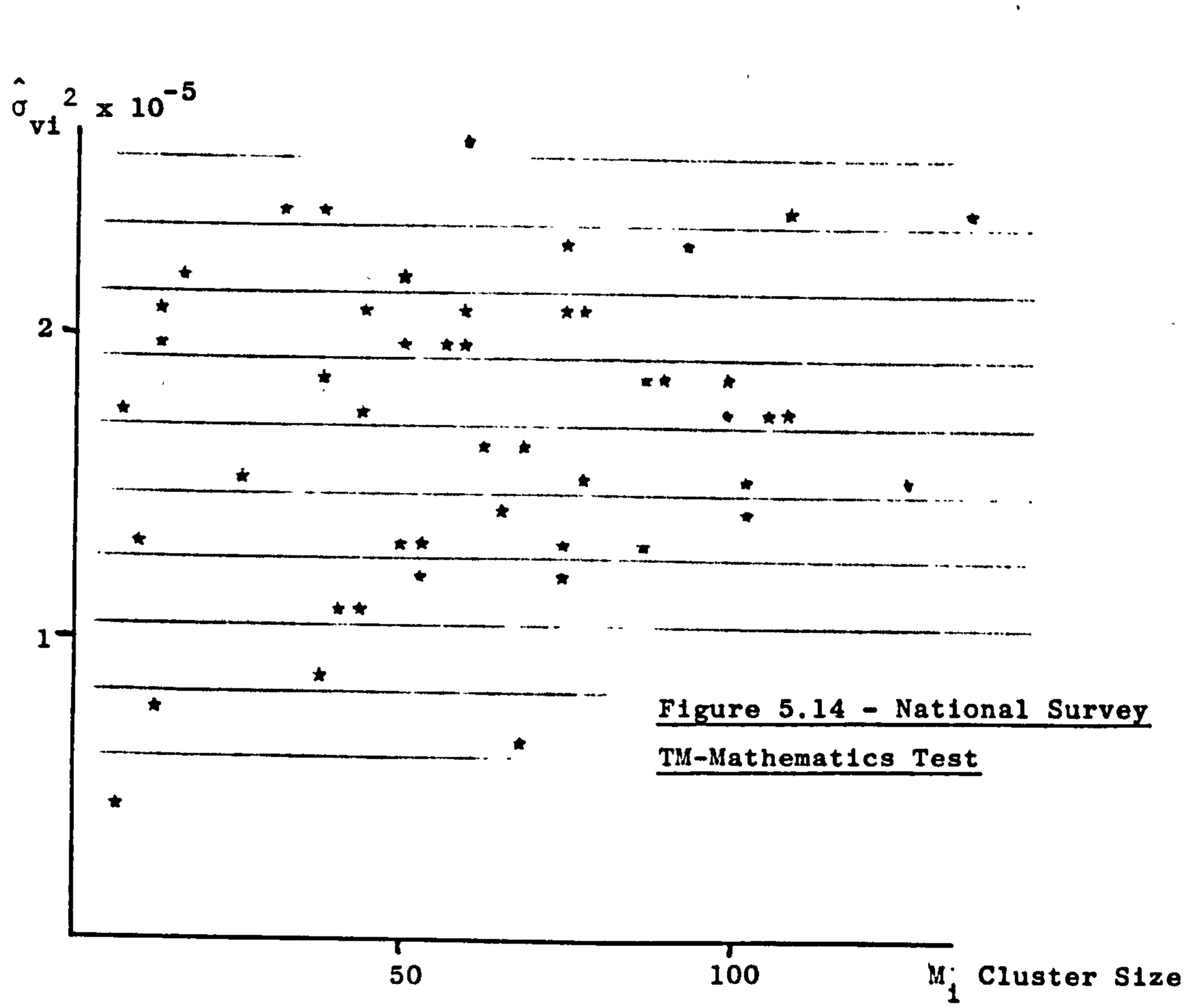


Figure 5.14 - National Survey
TM-Mathematics Test

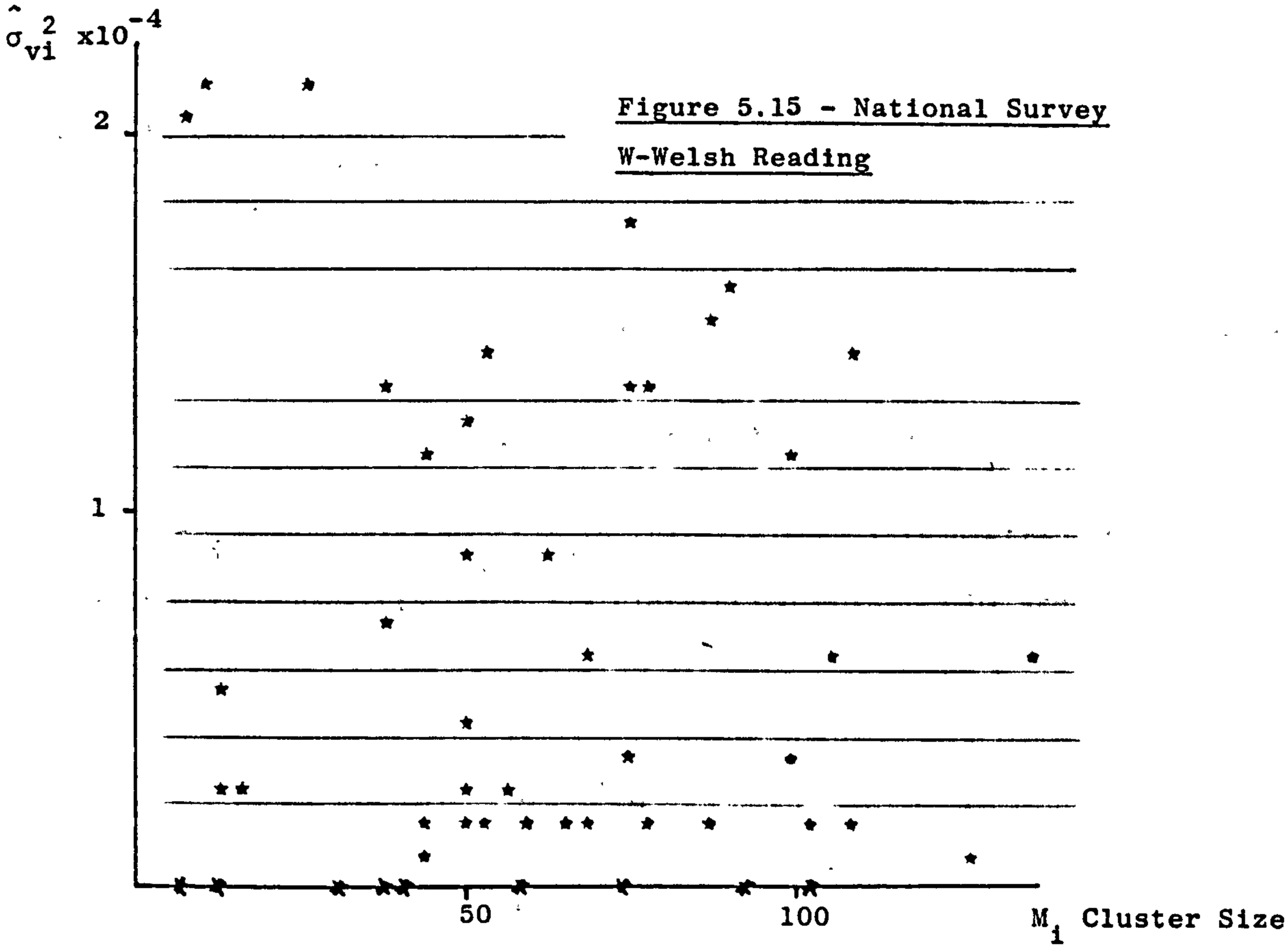


Figure 5.15 - National Survey W-Welsh Reading

As in Figures 5.4-5.6, there is little evidence here of the regression functions depending on M_i, let alone increasing to an asymptote. Here again Assumption A seems plausible.

We now compare the misspecification effects of T_{Ym} and $T_{Y\tilde{v}}$. We may restrict our comparison to that of $\tau_{Ym}(M_i)$ and $\tau_{Y\tilde{v}}(M_i)$ since the misspecification effects in (5.27) and (5.41) have the same form. It is helpful to consider some special cases separately.

Case 1: A holds, $\sigma_i^2 = \sigma_W^2$, $k_{3i} = k_{3W}$, i=1 ... N

If A holds

$$\tau_{Ym}(M) = \tau_{Ym}$$

$$\tau_{Y\tilde{v}}(M) = \tau_{Y\tilde{v}}$$

Algebraic Comparison of τ_{Ym} and $\tau_{Y\tilde{v}}$

From (5.30) and (5.45)

$$\tau_{Ym} = \sigma_B^2/\sigma^2 \tag{5.50}$$

$$\tau_{Y\tilde{v}} = (2\sigma_B^4 + k_{4B})/(2\sigma^4 + k_4) \tag{5.51}$$

If $k_{4B} = k_{4W} = 0$, as for example in the case when Y is distributed normally within clusters and μ_i is distributed normally between clusters, then

$$\tau_{Y\tilde{v}} = \tau_{Ym}^2$$

Hence $$\tau_{Y\tilde{v}} \leq \tau_{Ym}$$

Indeed, since τ_{Ym} is usually 'small' in most surveys, $\tau_{Y\tilde{v}}$ will be 'very small'. This is in accordance with Kish and Frankel's (1974) conjecture that deffs for complex statistics are smaller than deffs for means.

In the general non-normal case

$$\tau_{Y\tilde{v}} = \left(\frac{2 + K_B}{2 + \tau_{Ym}^2 K_B + (1-\tau_{Ym})^2 K_W}\right) \tau_{Ym}^2 \tag{5.52}$$

where K_B is the kurtosis of μ_i between clusters

K_W is the kurtosis within clusters

It is clear that $\tau_{Y\tilde{v}}$ will only be greater than τ_{Ym} in very exceptional cases when K_B is very positive and K_W is very negative (that is close to -2).

Graphic Comparison of τ_{Ym} and $\tau_{Y\tilde{v}}$

From (5.27) and (5.42) we may compare τ_{Ym} and $\tau_{Y\tilde{v}}$ by comparing plots of Y_{ij} against $Y_{ij'}$, $(j \neq j')$ with plots of $(Y_{ij} - \mu)^2$ against $(Y_{ij'} - \mu)^2$.

Example 5.6

Consider a population consisting of a mixture, according to equal proportions, of five types of cluster within which Y_{ij} is uniformly distributed on the intervals (0,2), (1,3), (2,4), (3,5) and (4,6) respectively. In Figure 5.16 Y_{ij} is plotted against $Y_{ij'}$, $(j \neq j')$ and within-cluster 95% probability squares are indicated. In Figure 5.17 $(Y_{ij} - \mu)^2$ is plotted against $(Y_{ij'} - \mu)^2$ (where $\mu=3$). In this figure within-cluster probability density contours are hyperbolae. The largest shape in Figure 5.17 corresponds to the two extreme clusters (0,2) and (4,6). These two clusters tend to dominate Figure 5.17 and to attenuate $\tau_{Y\tilde{v}}$.

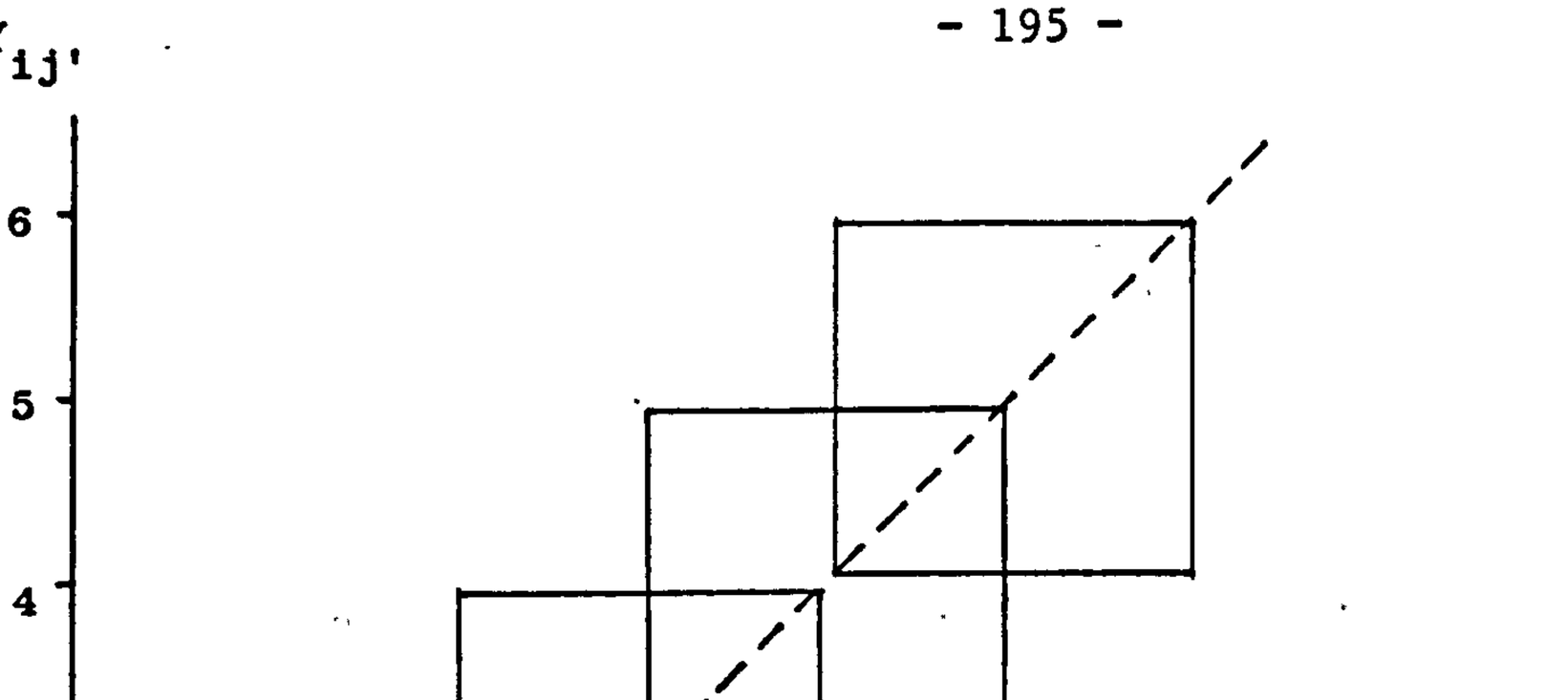


Figure 5.16 - Mixture of Uniform Distributions

95% Probability regions for Y, Y'

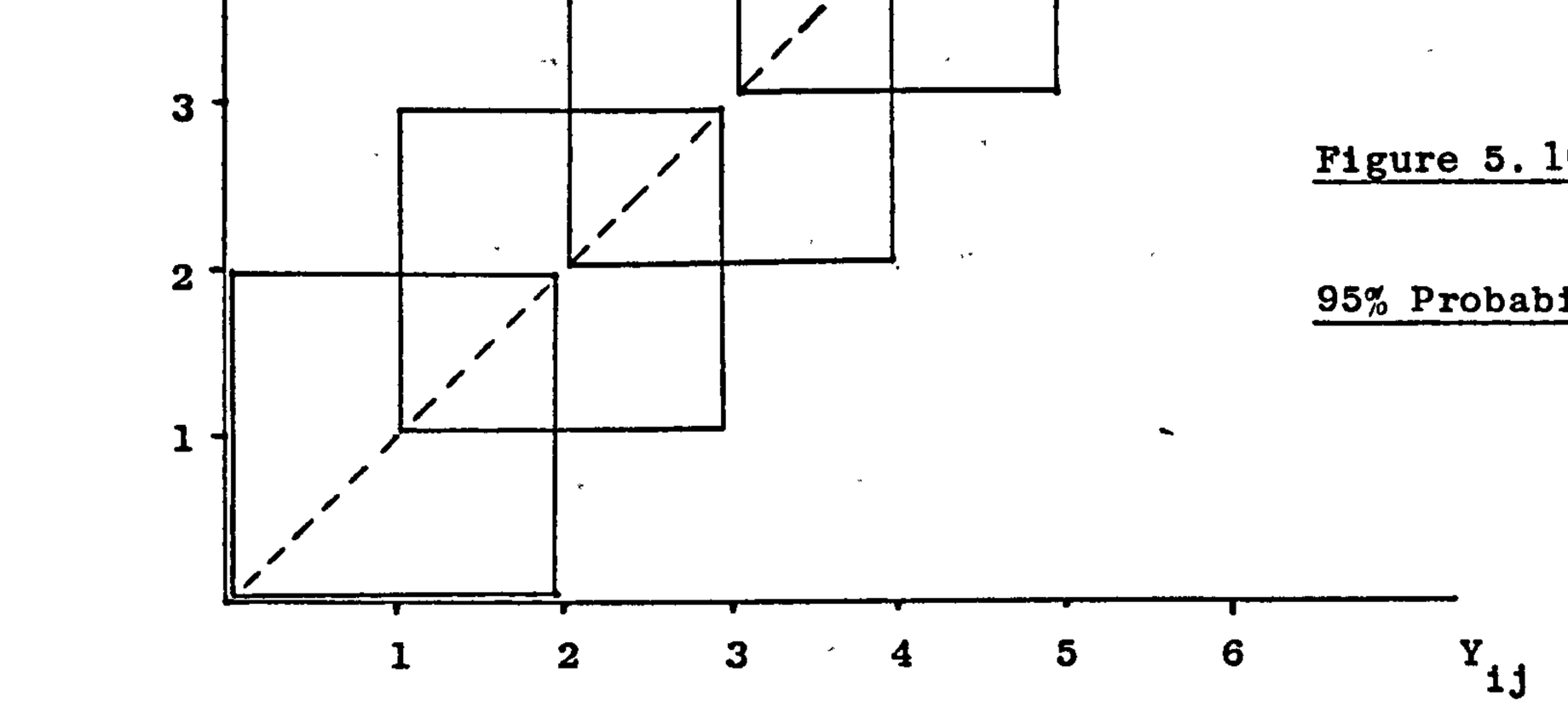


Figure 5.17 - Mixture of Uniform Distributions

95% Probability regions for $(Y - \mu)^2$, $(Y' - \mu)^2$

The two Figures graphically illustrate how τ_{Ym} may be higher than $\tau_{Y\tilde{v}}$.

A similar comparison would be obtained if the within cluster distribution were normal. In this case the constant probability contours in Figure 5.16 would be circles of constant diameter and the contours in Figure 5.17 would be approximately hyperbolae (the $(Y_{ij} - \mu)^2$ have non-central chi-squared distributions), the tails of which tend to attenuate $\tau_{Y\tilde{v}}$.

Case 2: A holds, $\mu_i = \mu$, i=1 ... N

As in Case 1, if A holds,

$$\tau_{Ym}(M_i) = \tau_{Ym}$$

$$\tau_{Y\tilde{v}}(M_i) = \tau_{Y\tilde{v}}$$

Also in this case, from (5.30),

$$\tau_{Ym} = 0$$

Algebraic Comparison of τ_{Ym} and $\tau_{Y\tilde{v}}$

From (5.45)

$$\tau_{Y\tilde{v}} = \gamma/(2\sigma_W^4 + k_{4W} + 3\gamma) \qquad (5.53)$$

Hence if $\gamma \neq 0$ then $\tau_{Y\tilde{v}} > \tau_{Ym}$ contradicting Kish and Frankel's (1974) conjecture that deffs for complex statistics are not greater than deffs for means.

We may interpret the expression for $\tau_{Y\tilde{v}}$ in (5.53) by recalling that τ_{Ym} may be viewed as a measure of homogeneity of cluster means μ_i (if A holds)

$$\tau_{Ym} = \text{corr}_I(Y_{ij}, Y_{ij'}) \quad (j \neq j')$$

An analogous measure of homogeneity of cluster variances σ_i^2 may be defined as

$$\tau_{Yv2} = \text{corr}_I\left[(Y_{ij} - \mu_i)^2, (Y_{ij'} - \mu_i)^2\right] \qquad (j=j')$$

$$= \text{var}_I(\sigma_i^2)/\text{var}_I\left[(Y_{ij} - \mu_i)^2\right]$$

$$= \gamma/(2\sigma_W^4 + k_{4W} + 3\gamma) \qquad (5.54)$$

again assuming that A holds. We see from (5.53) and (5.54) that in Case 2

$$\tau_{Y\tilde{v}} = \tau_{Yv2}$$

Graphic Comparison of τ_{Ym} and $\tau_{Y\tilde{v}}$

As for Case 1 we illustrate how τ_{Ym} and $\tau_{Y\tilde{v}}$ may be compared graphically, by example.

Example 5.7

Consider a population consisting of a mixture (according to arbitrary proportions) of three types of cluster within which Y_{ij} is uniformly distributed on the intervals (0,6), (1,5) and (2,4) respectively. In Figure 5.18 Y_{ij} is plotted against $Y_{ij'}$, $(j \neq j')$ and the concentric within-cluster 95% probability squares are indicated. We note that Y_{ij} and $Y_{ij'}$ are uncorrelated not only within clusters but also across the whole population as we would expect if $\tau_{Ym} = 0$. In Figure 5.19 $(Y_{ij} - \mu)^2$ is plotted against $(Y_{ij'} - \mu)^2$ (where $\mu=3$). As in Figure 5.17 within-cluster probability contours are hyperbolae and 95% probability regions are indicated. In Figure 5.19 there is a

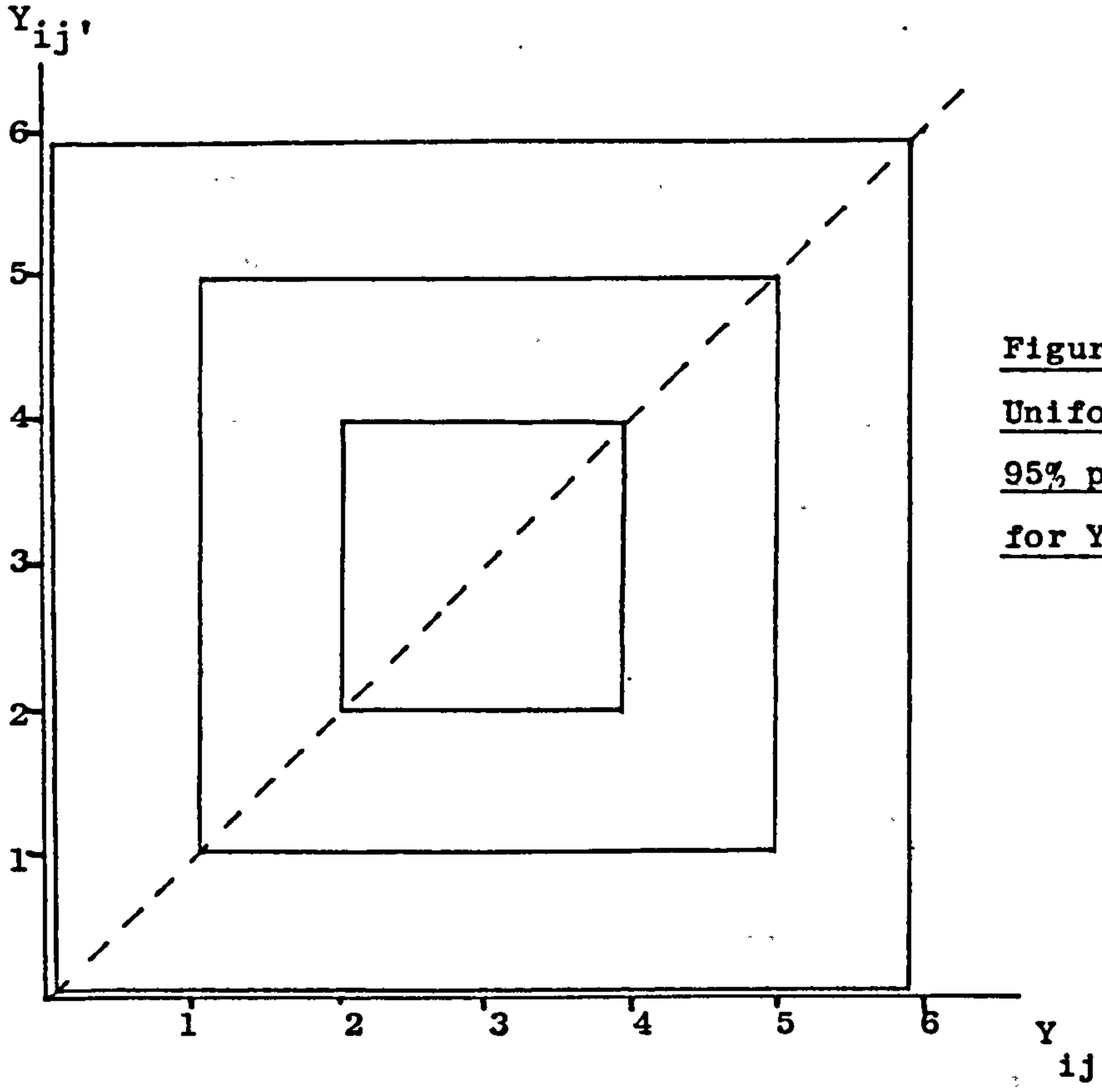


Figure 5.18 - Mixture of Uniform Distributions 95% probability regions for Y, Y'

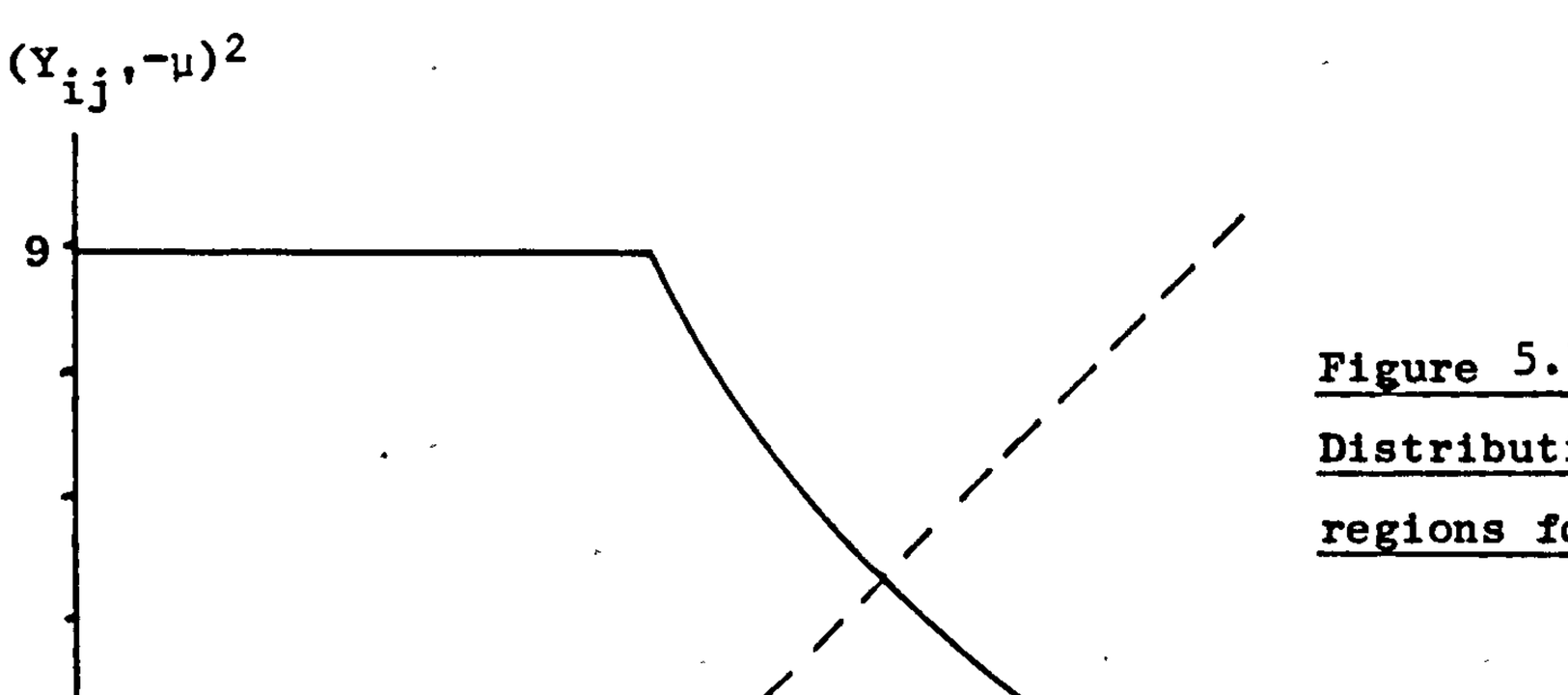


Figure 5.19 - Mixture of Uniform Distributions - 95% probability regions for $(Y - \mu)^2$, $(Y' - \mu)^2$

positive, though low, correlation between $(Y_{ij} - \mu)^2$ and $(Y_{ij'} - \mu)^2$ indicating a positive value of $\tau_{Y\tilde{v}}$.

In order to recognise the sign of the population correlation in diagrams such as Figure 5.19 we note that if W denotes the within cluster distribution and B denotes the between-cluster distribution and if Z and Z' are two random variables such that

$$cov_W(Z,Z') = 0$$

then

$$cov(Z,Z') = cov_{WB}(Z,Z')$$

$$= cov_B(E_W(Z),\ E_W(Z'))$$

In Figure 5.19 the pairs $(E_W(Z),\ E_W(Z')$ are the cluster centroids and are denoted by *. Clearly the centroids are positively correlated in this figure and so $\tau_{Y\tilde{v}}$ is positive.

Example 5.18

Consider a population consisting of a mixture of three types of cluster within which Y_{ij} is normally distributed with zero means and standard deviations $\sigma = 1,2,3$ respectively. In Figure 5.20 Y_{ij} is plotted against $Y_{ij'}$, $(j \neq j')$ and the concentric 95% probability circles are indicated (with radii $\sqrt{-2\ \log(.05)}\ \sigma = 2.45\sigma$). Y_{ij} and $Y_{ij'}$ are uncorrelated both within clusters and across the whole population. In Figure 5.21 $Z = (Y_{ij} - \mu)^2$ is plotted against $Z' = (Y_{ij'} - \mu)^2$ (where $\mu = 0$). Constant within-cluster probability density contours are defined by the equation

$$\log(ZZ'/\sigma^4) + (Z + Z')/\sigma^2 = C$$

for different constants C, since Z/σ^2 and Z'/σ^2 are independently

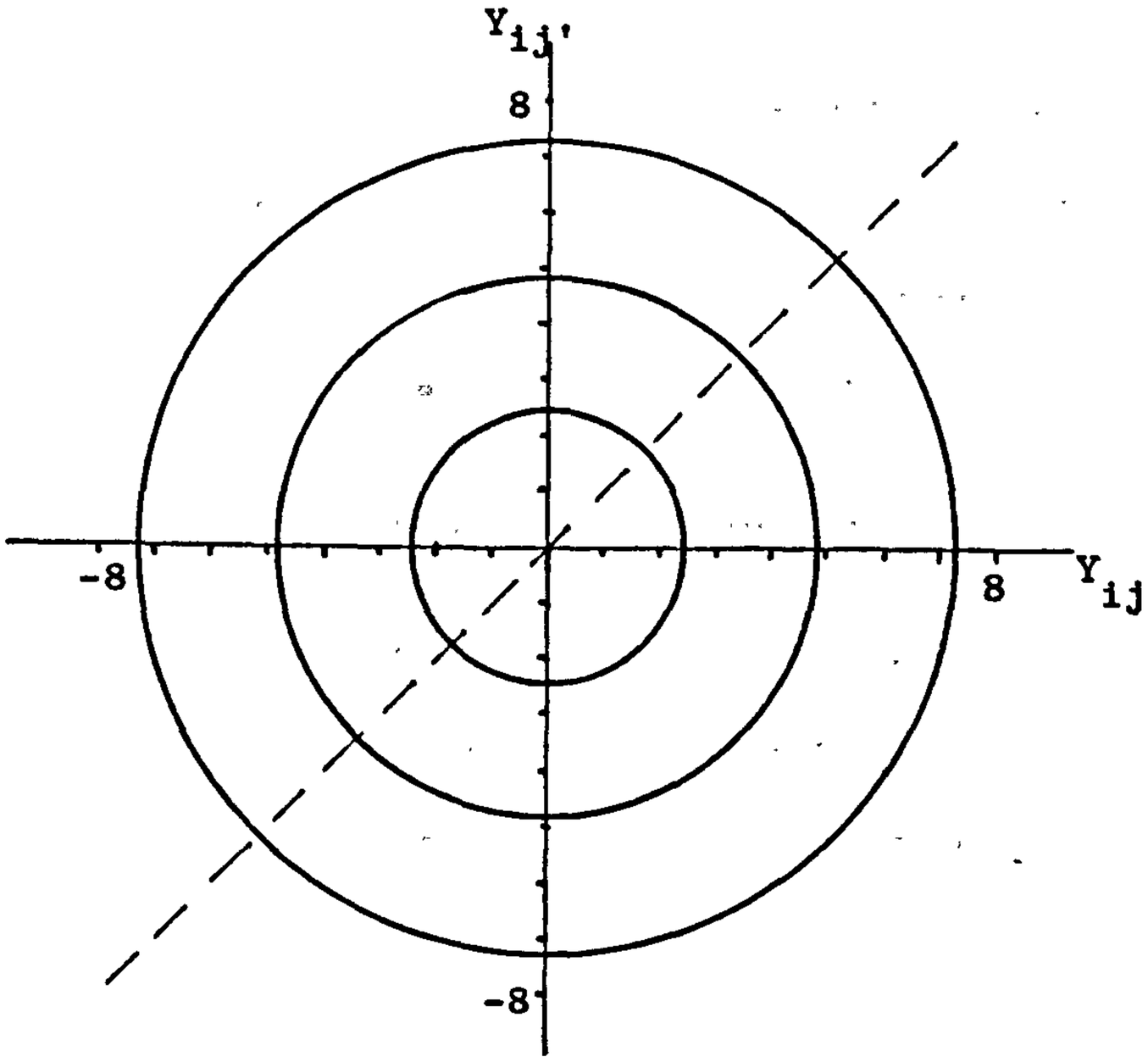


Figure 5.20 - Mixture of Normal Distributions 95% confidence regions for Y, Y'

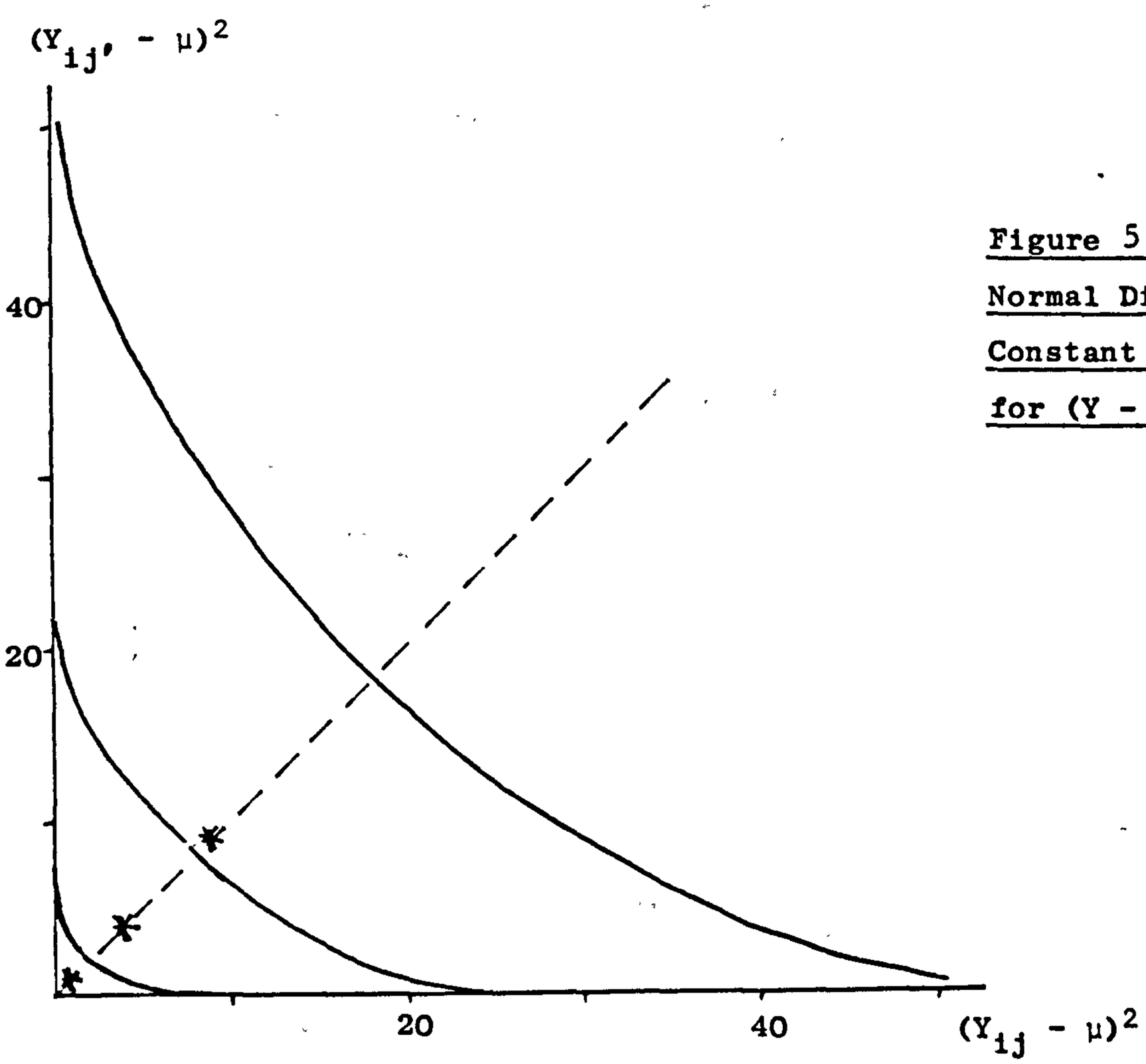


Figure 5.21 - Mixture of Normal Distributions - Constant Probability Regions for $(Y - \mu)^2$, $(Y' - \mu)^2$

distributed within clusters as chi-squared random variables with one degree of freedom. It appears to be a difficult mathematical problem to find the value of C corresponding to a 95% probability region. Instead we mark the contours in Figure 5.21 defined by

$$\log(ZZ'/\sigma^4) + (Z + Z')/\sigma^2 = \log(4) + 4$$

which pass through the points $(2\sigma^2, 2\sigma^2)$ and which define regions all having the same within-cluster probability content. As in Example 5.2 the position of the centroids indicates a positive (but low) value of $\tau_{Y\tilde{v}}$ compared with a zero value of τ_{Ym}.

Case 3: A holds

We now consider the general case when A holds. As in cases 1 and 2, $\tau_{Ym}(M_i) = \tau_{Ym}$, $\tau_{Y\tilde{v}}(M_i) = \tau_{Y\tilde{v}}$.

Algebraic Comparisons of τ_{Ym} and $\tau_{Y\tilde{v}}$

From (5.30)

$$\tau_{Ym} = \sigma_B^2/\sigma^2$$

In this case $\tau_{Y\tilde{v}}$ is a combination of the expressions (5.51) and (5.53) obtained in the last two cases.

Let $\tau_{Yv1} = (2\sigma_B^4 + k_{4B})/(2\sigma^4 + k_{4W} + k_{4B})$

Let $r_v = \text{corr}_I[(\mu_i - \mu)^2, \sigma_i^2]$

$= c_1(2\sigma_B^4 + k_{4B})^{-\frac{1}{2}} \gamma^{-\frac{1}{2}}$

Then from (5.45)

$$\tau_{Y\tilde{v}} = w_{v1}\,\tau_{Yv1} + 2r_v(w_{v1}\,\tau_{Yv1} w_{v2}\,\tau_{Yv2})^{\frac{1}{2}} + w_{v2}\,\tau_{Yv2} \qquad (5.55)$$

where $w_{v1} = (2\sigma^4 + k_{4W} + k_{4B})/(2\sigma^4 + k_4)$

$w_{v2} = (2\sigma_W^4 + k_{4W} + 3\gamma)/(2\sigma^4 + k_4)$

and τ_{Yv2} is given in (5.54).

Note that in Case 1

$$w_{v1} = 1 \ , \ \tau_{Yv2} = 0 \ , \ \tau_{Y\tilde{v}} = \tau_{Yv1}$$

and in Case 2

$$\tau_{Yv1} = 0 \ , \ w_{v2} = 1 \ , \ \tau_{Y\tilde{v}} = \tau_{Yv2}$$

In Table 5.2 we give some parameter estimates based on 1975 Family Expenditure Survey data (for the method of estimation see the Appendix). The FES sample design is described in Kemsley (1969). To summarise, 1782 adminstrative areas of Great Britain were divided into 168 strata using regional, area-type and economic stratification factors. Within each stratum a single adminstrative area was selected as a psu by PPS and retained in the sample for four quarters according to a rotation scheme which replaced 42(= 168/4) psu's each quarter. Within each psu a new ward or group of parishes) was selected as a ssu by PPS at each quarter. Within each ssu, 16 addresses (households) were selected by srs. Hence the overal design was epsem.

We make the simplifying assumption that each cluster consists of all households selected within the stratum in the given year, 1975. This yields 7054 households (out of 168 x 16 x 4 = 10752 possible households) divided into n = 168 clusters of approximately equal size ($\bar{m}$ = 42.0, m* = 42.6). We thus ignore stratification, differences between psu's within strata, differences between quarters and non-response. These simplifications mean that the m_i and n are large enough for the sampling errors to be small. The variables considered were:

V1 : normal gross income

V2 : expenditure on food

V3 : total expenditure

Logarithms of each variable were taken to remove skewness. The cluster sizes, M_i, vary from about 3000 in rural areas to about 250,000 in the GLC. We may therefore expect Assumption B to be invalid for the variables of interest although given Lemma 5.11 and the self-weighting design we might hope that this does not matter too much.

From Table 5.2 we see that

$$\tau_{Y\tilde{v}} \doteq \tau_{Yv2}$$

The fact that $\tau_{Y\tilde{v}}$ is mainly a function of the dispersion in cluster variances (as measured by τ_{Yv2}) is to be expected. For, when the effect of clustering is not 'too great', the population moments σ^2 and k_4 are approximately equal to the within-cluster moments σ_w^2 and k_{4w} and so w_{v1} and w_{v2} will be approximately unity (as above). Furthermore, unless the within or between cluster kurtosis is severe, τ_{Yv1} will be approximately equal to τ_{Ym}^2 which will be very small (as above). Hence we might expect $\tau_{Y\tilde{v}}$ to be dominated by τ_{Yv2}.

Table 5.2 - Parameter Estimates for the Family Expenditure Survey

Variable Y	$\hat{\tau}_{Ym}$	$\hat{w}_{v1}$	$\hat{\tau}_{Yv1}$	$\hat{r}_v$	$\hat{w}_{v2}$	$\hat{\tau}_{Yv2}$	$\hat{\tau}_{Y\tilde{v}}$	$\hat{meff}(T_{Ym})$	$\hat{meff}(T_{Y\tilde{v}})$
log(V1)	.032	.993	.001	-.301	.941	.004	.0033	2.325	1.137
log(V2)	.016	.921	.000	-.026	.971	.027	.0267	1.675	2.112
log(V3)	.031	.981	.001	-.112	.942	.007	.0072	2.284	1.300

It is clear from the above that $\tau_{Y\tilde{v}}$ depends fundamentally on the joint distribution of μ_i and σ_i^2. By means of five examples we now suggest how various relationships between μ_i and σ_i^2 might arise.

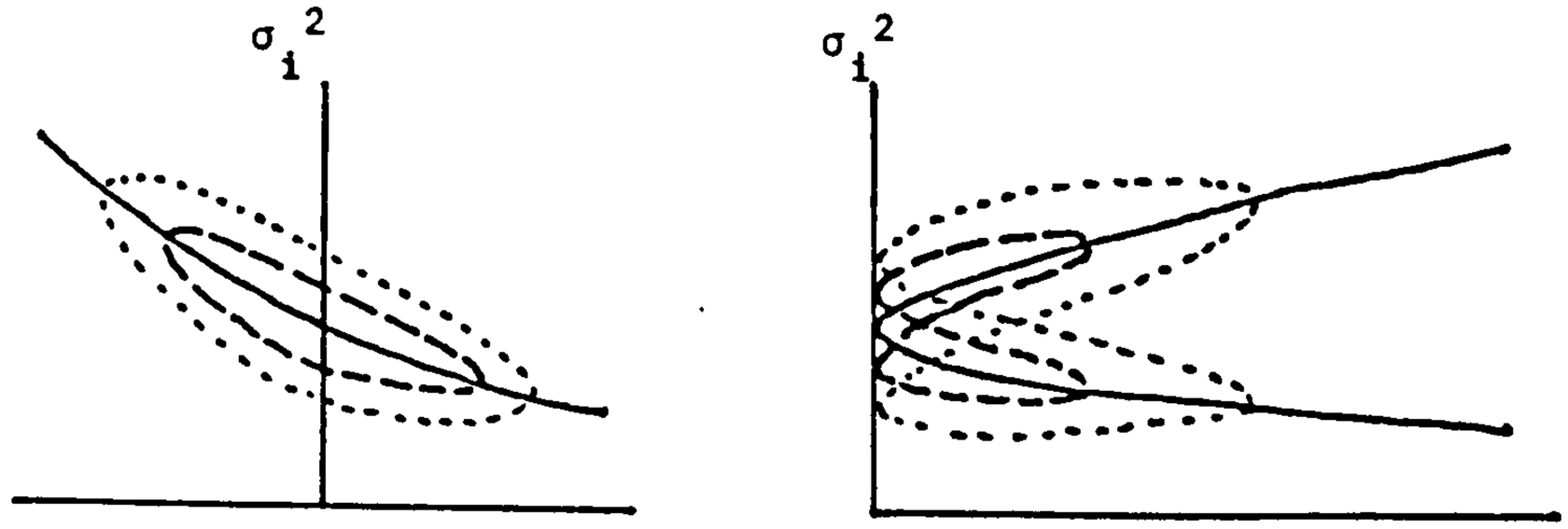
(i) Y_{ij} has a Poisson within-cluster distribution with parameter λ_i.

Then $\mu_i = \sigma_i^2 = \lambda_i$.

(ii) Y_{ij} has a lognormal within-cluster distribution which varies only in its scale parameter between clusters. Then $\sigma_i^2 \propto \mu_i^2$.

Scatter diagrams of $\hat{\mu}_i$ ($= \bar{y}_i$) against $\hat{\sigma}_i^2$ (see 5.32) for two variables, which we might expect to be lognormally distributed, are given in Figures 5.22 and 5.23. In both examples (i) and (ii) and in these scatter diagrams μ_i and σ_i^2 are related monotonically. In this case $(\mu_i - \mu)^2$ and σ_i^2 will be related non-monotonically and we might expect r_v to be small. This situation is illustrated in Figure 5.24.

Figure 5.24



We now give three examples where μ_i and σ_i^2 are not related monotonically.

(iii) In Figure 5.25 $\hat{\mu}_i$ is plotted against $\hat{\sigma}_i^2$ for the variable

Figure 5.22 - General Household Survey

Y = Gross Value of Property

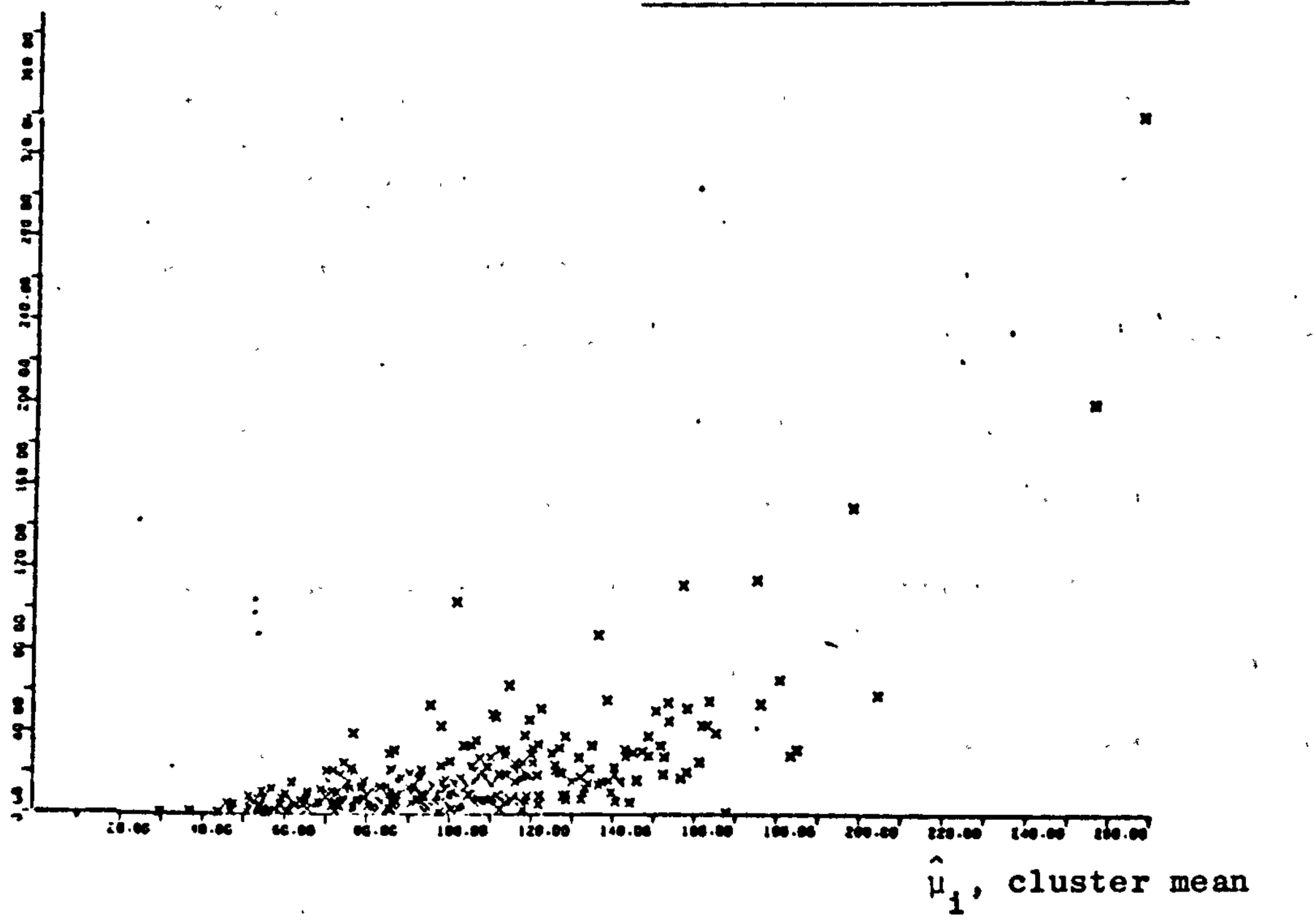
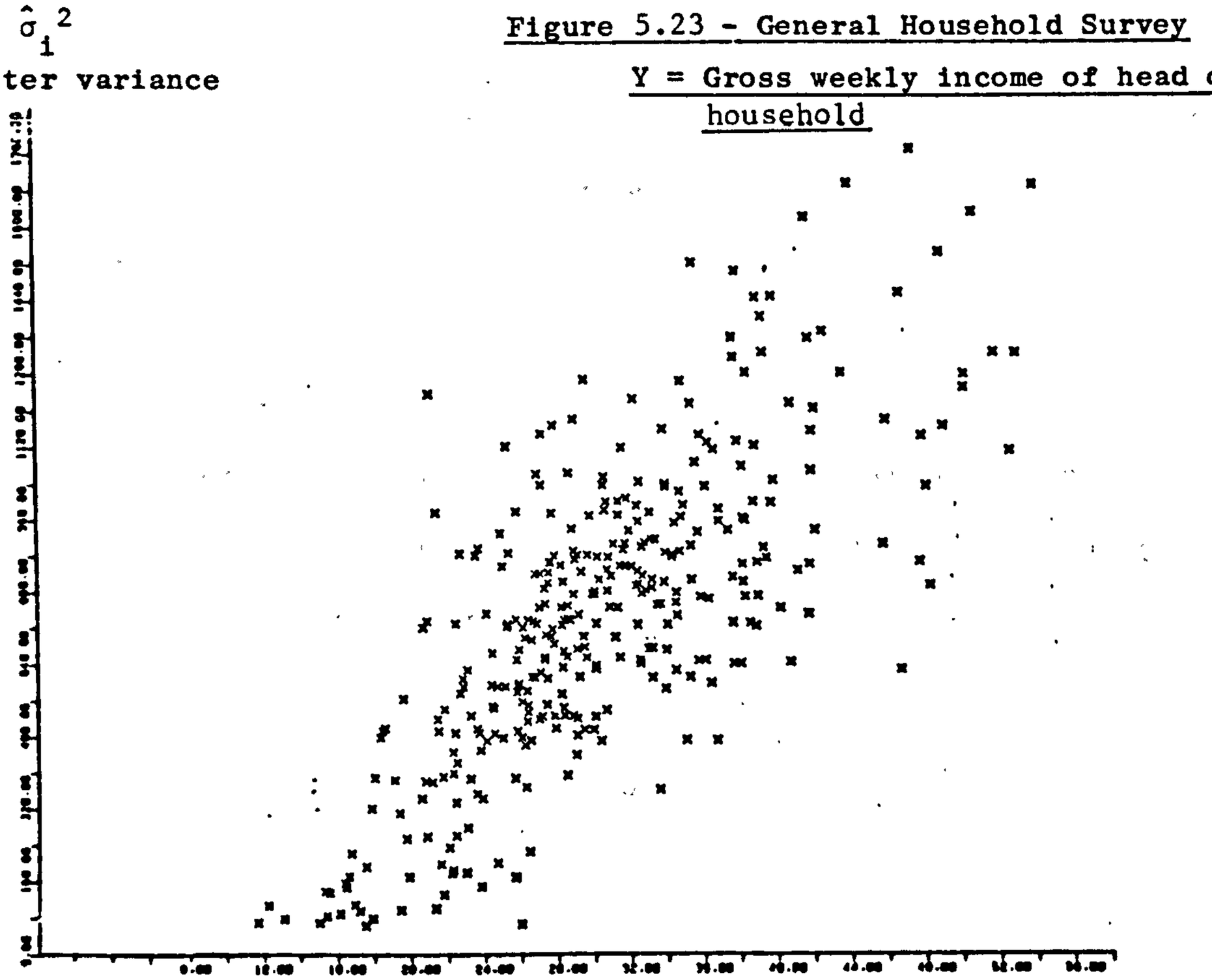


Figure 5.23 - General Household Survey

Y = Gross weekly income of head of household



(iii) (Continued)

TM (mathematics test) from the National Survey of Attainment data. Two groups of clusters (schools) are evident. A simple model for such data is as follows.

TM scores are distributed in the population according to the distribution F defined on the interval $[a,b]$ (a=10, b=80).

For schools in group 1, TM scores are distributed according to F truncated above by c_i ($\lesssim$ b). Hence μ_i and σ_i^2 increase as c_i increases and so σ_i^2 increases as μ_i increases in group 1.

For schools in group 2, TM scores are distributed according to F truncated below by c_i ($\gtrsim$ a). Hence μ_i increases but σ_i^2 decreases as c_i increases and so σ_i^2 decreases as μ_i increases.

In Figure 5.26 $\hat{\sigma}_i^2$ is plotted against $(\hat{\mu}_i - \hat{\mu})^2$ and the relationship is roughly monotonic since $\hat{\mu}$ lies between the $\hat{\mu}_i$ values for the two groups of schools. In this case we would expect r_v to be non-negligible.

(iv) In Figure 5.27 $\hat{\mu}_i$ is plotted against $\hat{\sigma}_i^2$ for the variable W (Welsh reading) from the National Survey of Attainment data.

A very simplified model is as follows.

Individuals in the population are either 'Welsh speakers', who always score W=b, or 'Non-Welsh speakers' who always score W=a.

If the proportion of Welsh speakers in the i^{th} cluster is p_i then

$$\mu_i = (1-p_i)a + p_i\, b$$

$$\sigma_i^2 = (b-a)^2 p_i(1-p_i)$$

$$= (\mu_i - a)/(b-\mu_i)$$

In Figure 5.27 a line corresponding to a=4, b=24 is drawn.

TEXT BOUND INTO

THE SPINE

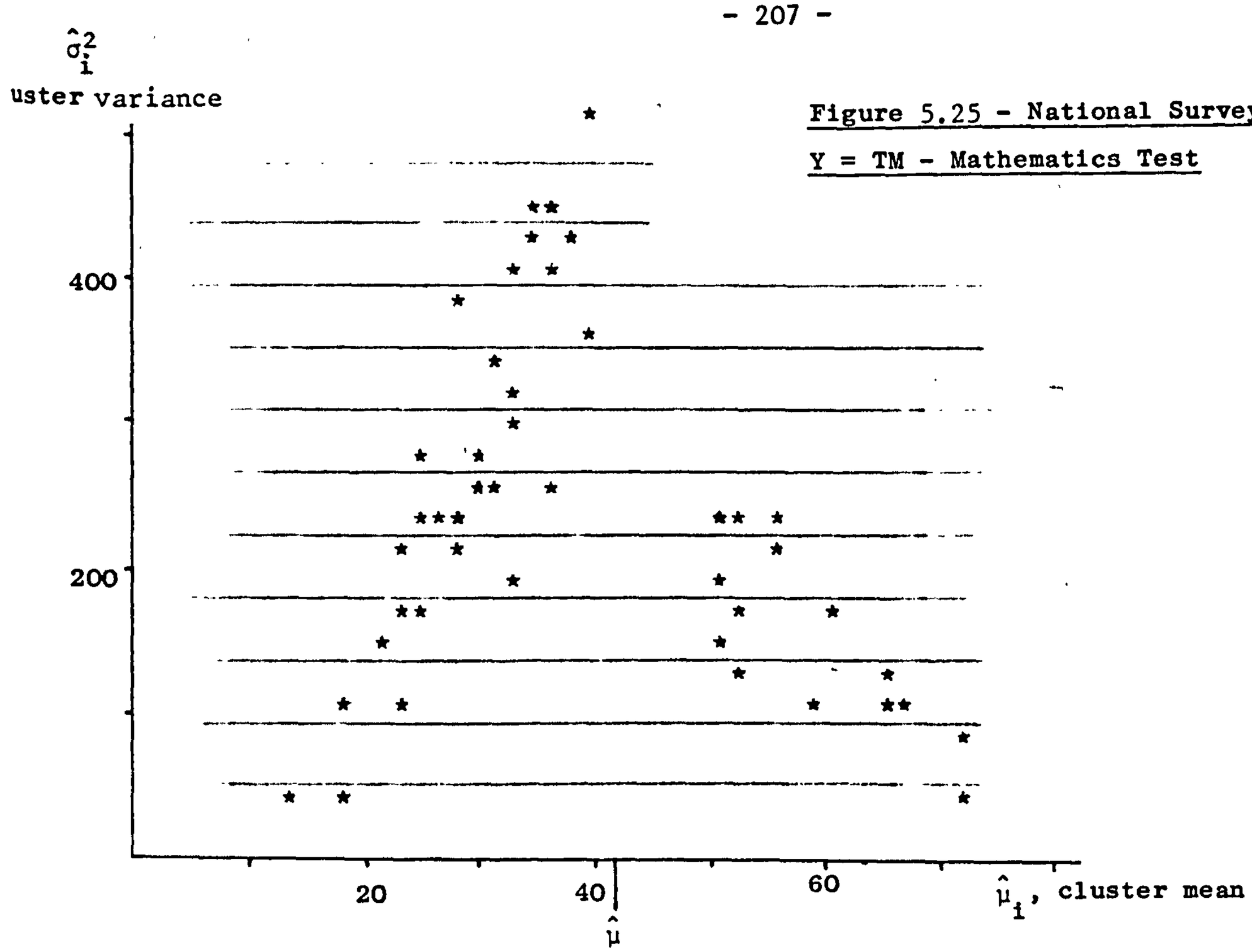


Figure 5.25 - National Survey

Y = TM - Mathematics Test

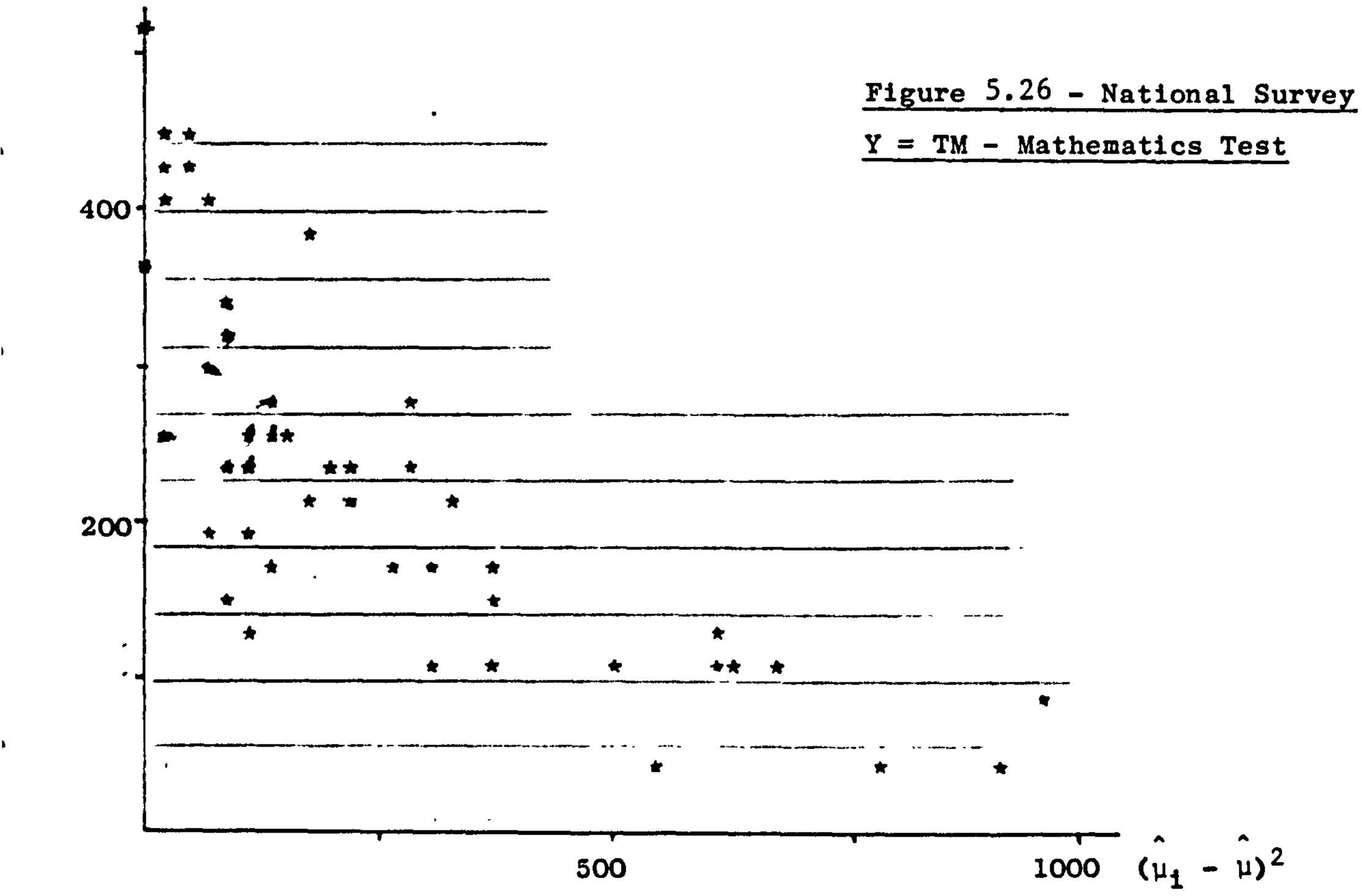


Figure 5.26 - National Survey

Y = TM - Mathematics Test

Figure 5.27. - National Survey

Y = W - Welsh Language

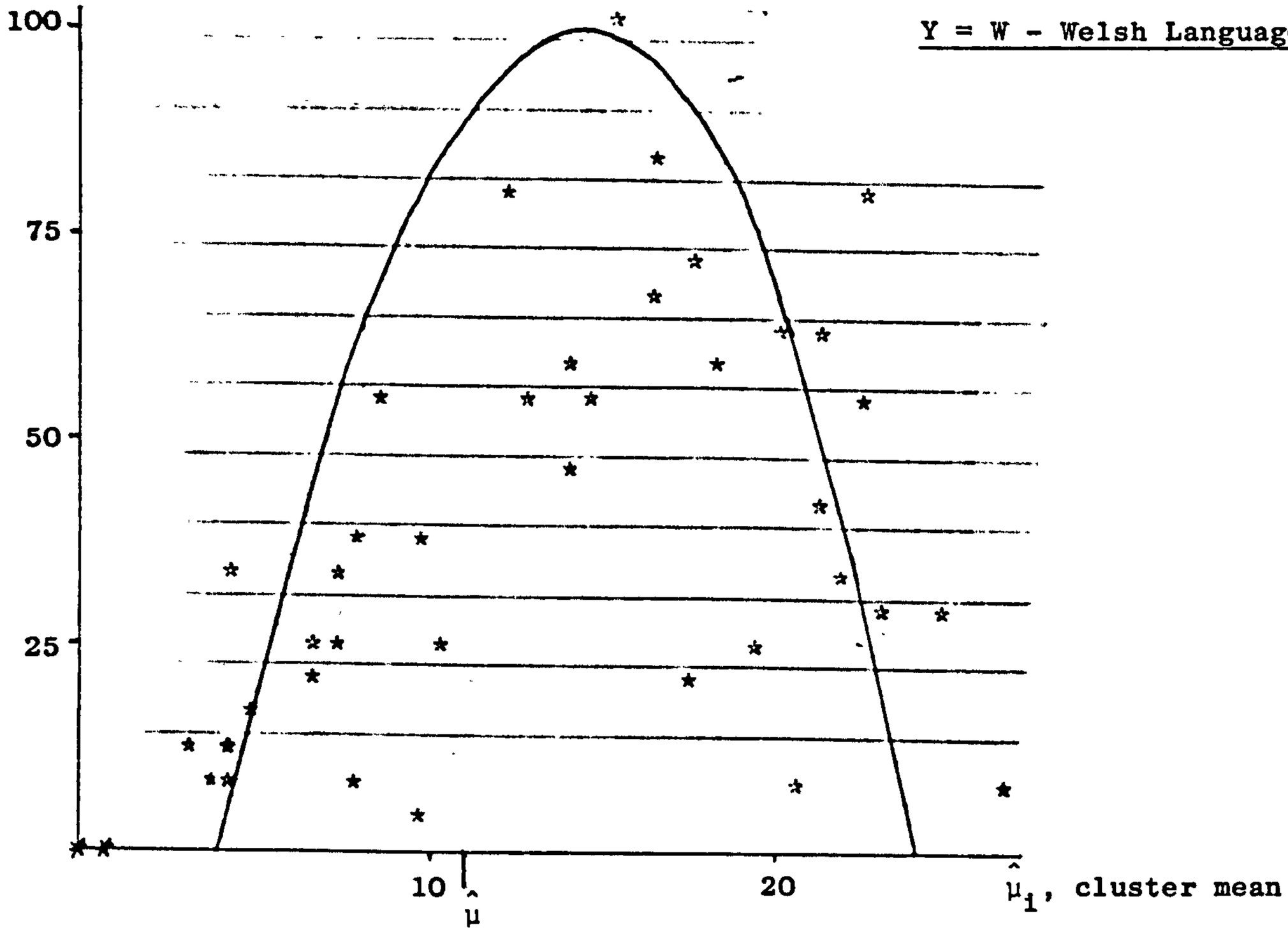
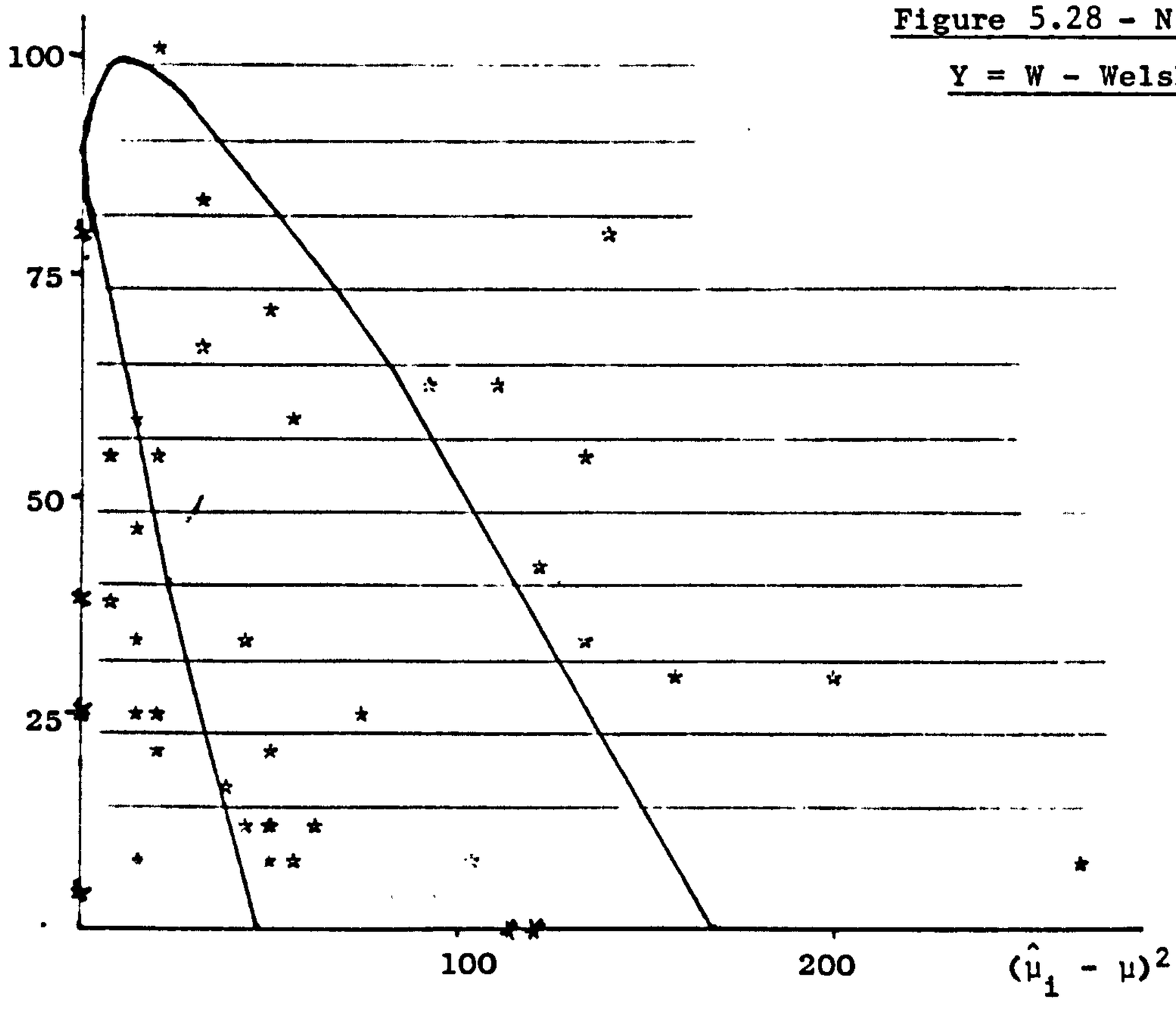


Figure 5.28 - National Survey

Y = W - Welsh Language



(iv) (Continued)

Since the distribution of p_i's is not symmetric (there are more non-Welsh speakers than Welsh-speakers) the relationship between σ_i^2 and $(\mu_i - \mu)^2$ as depicted in Figure 5.28 is non-monotonic and is not unlike Figure 5.24.

(v) In (iv) the within-cluster distribution of W is a linear transformation of a Bernoulli distribution. This may be generalised to the situation where the population is a mixture of two groups with means, a and b, and common variances c^2. If the proportion of group 2 in the i^{th} cluster is p_i then

$$\mu_i = (1-p_i)a + p_i\, b$$

$$\sigma_i^2 = c^2 + (b-a)^2 p_i(1-p_i)$$

$$= c^2 + (\mu_i - a)(b-\mu_i)$$

If $E(p_i) = 0.5$ then

$$\mu = (a+b)/2$$

and $$\sigma_i^2 = c^2 + (b-a)^2/4 - (\mu_i-\mu)^2$$

So σ_i^2 is linearly related to $(\mu_i-\mu)^2$

Moreover $\sigma_i^2 + (\mu_i-\mu)^2$ is constant and so $\tau_{Y\tilde{v}} = 0$ (see Lemma 5.17) even though $\tau_{Ym} \neq 0$. In general if $E(p_i) \neq 0.5$ then the relationship between σ_i^2 and $(\mu_i-\mu)^2$ will not be monotonic as in Figure 5.28.

Geometric Comparison of τ_{Ym} and $\tau_{Y\tilde{v}}$

We indicate how τ_{Ym} and $\tau_{Y\tilde{v}}$ may depend on the joint distribution of μ_i and σ_i^2 by example.

Example 5.9

Consider two populations

Population 1 - a mixture (in equal proportions) of six clusters within which Y_{ij} is uniformly distributed on the intervals $(-\sqrt{3}, -\sqrt{2})$, $(-\sqrt{2}, -1)$, $(-1, 0)$, $(0, 1)$, $(1, \sqrt{2})$, $(\sqrt{2}, \sqrt{3})$ respectively.

Population 2 - a mixture (in equal proportions) of six clusters within which Y_{ij} is uniformly distributed on the intervals $(-\sqrt{3}, -\sqrt{3} + 1)$, $(-\sqrt{3} + 1, -\sqrt{3} + \sqrt{2})$, $(-\sqrt{3} + \sqrt{2}, 0)$, $(0, \sqrt{3} - \sqrt{2})$, $(\sqrt{3} - \sqrt{2}, \sqrt{3} - 1)$, $(\sqrt{3} - 1, \sqrt{3})$ respectively.

The marginal distributions of μ_i and σ_i^2 are the same in both populations. Hence τ_{Ym} is the same in both populations, as can be verified graphically by comparing Figures 5.29 and 5.30. However, σ_i^2 and $(\mu_i - \mu)^2$ are negatively correlated in population 1 whereas they are positively correlated in population 2. By inspecting Figures 5.31 and 5.32 it appears that $\tau_{Y\tilde{v}}$ is greater in population 1 than in population 2. This is somewhat unexpected since from (5.55) we might expect $\tau_{Y\tilde{v}}$ to be greater when r_v is positive.

To investigate the dependence of $\tau_{Y\tilde{v}}$ on r_v, consider the special case $c_2 = k_{YB} = k_{YW} = 0$. Then from (5.45)

$$\tau_{Y\tilde{v}} = \frac{2\sigma_B^4 + 2c_1 + \gamma}{2\sigma^4 + 6c_1 + 3\gamma}$$

$$= \frac{1}{3} + \frac{2\sigma^4(\tau_{Ym}^2 - \frac{1}{3})}{2\sigma^4 + 6c_1 + 3\gamma}$$

Now σ^2, τ_{Ym} and γ are held constant across our populations. So if $\tau_{Ym} > 1/\sqrt{3} = .58$ $\tau_{Y\tilde{v}}$ decreases as c_1 or r_v increases, whereas if

Figure 5.29 - Population 1
95% probability regions
for Y, Y'

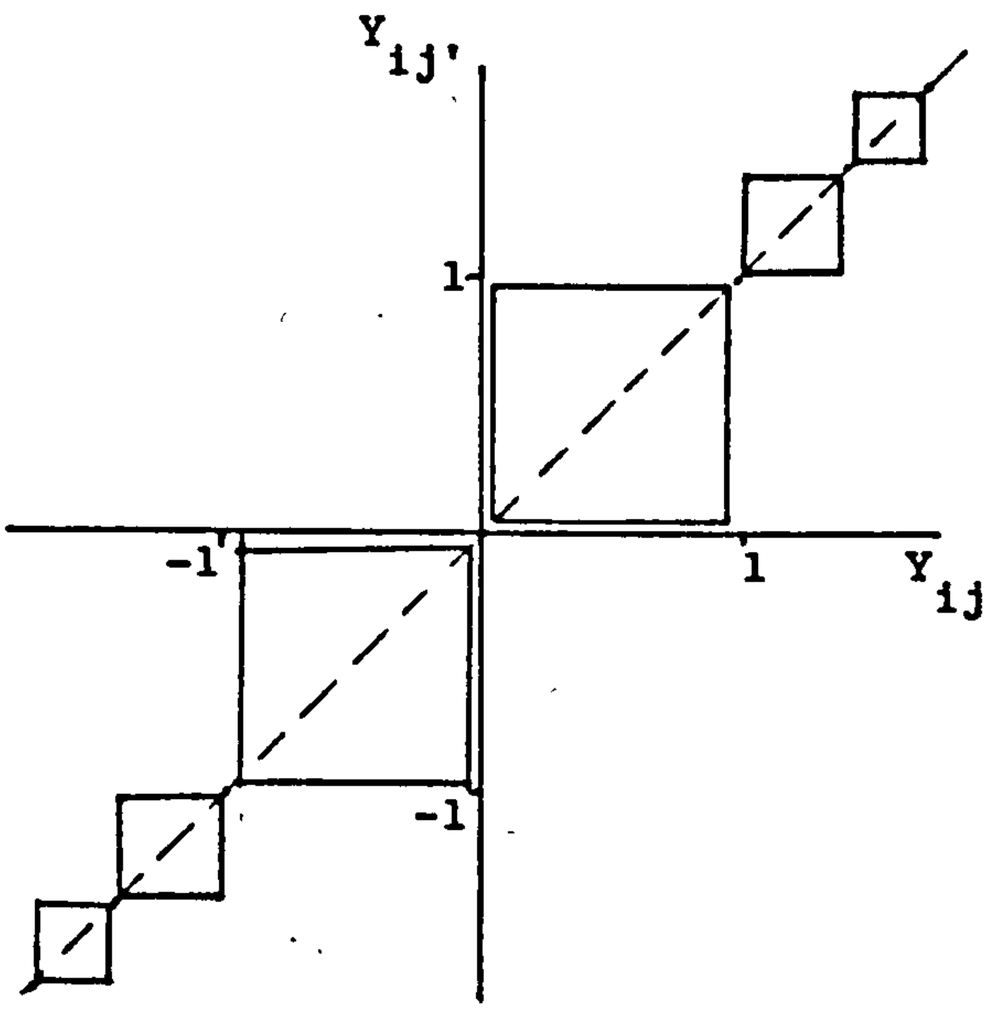


Figure 5.30 - Population 2
95% probability regions for Y, Y'

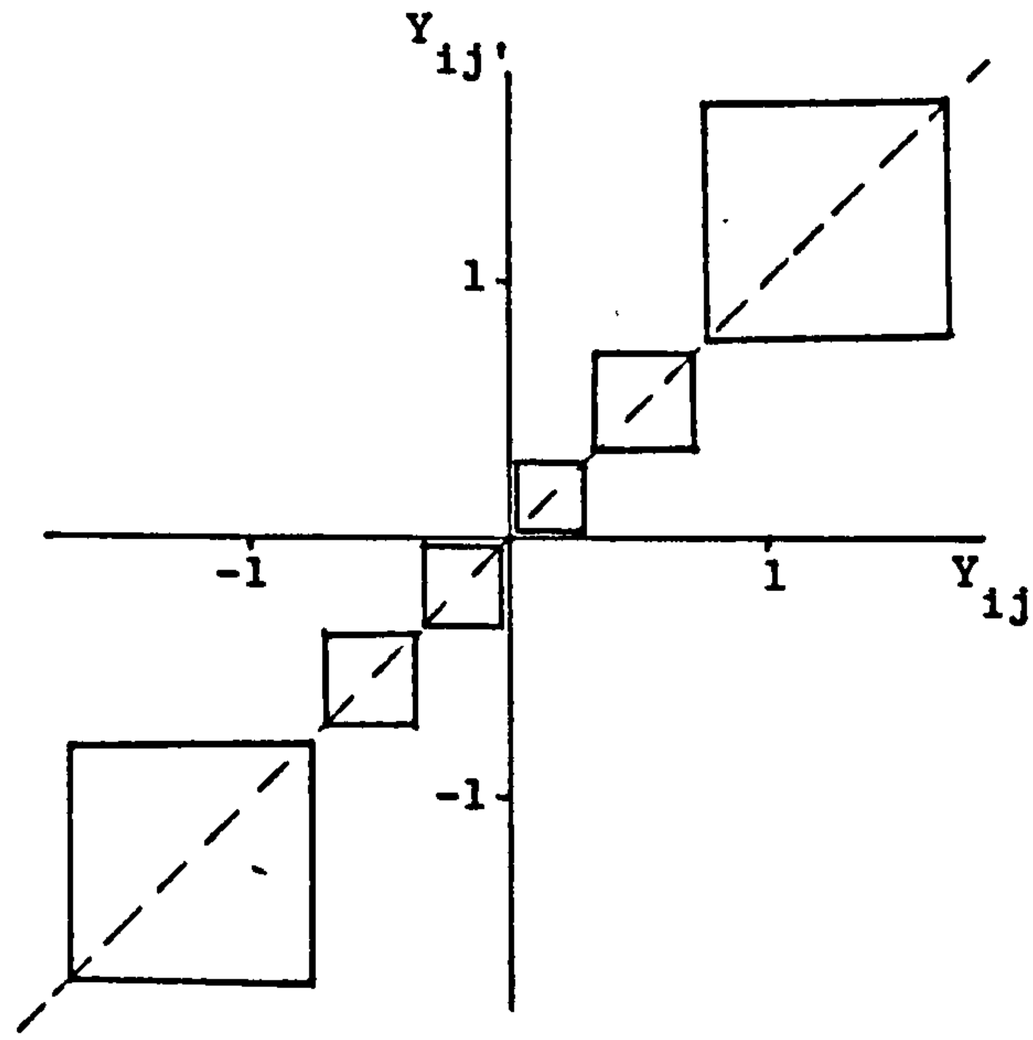


Figure 5.31 - Population 1
95% probability regions for
$(Y - \mu)^2$, $(Y' - \mu)^2$

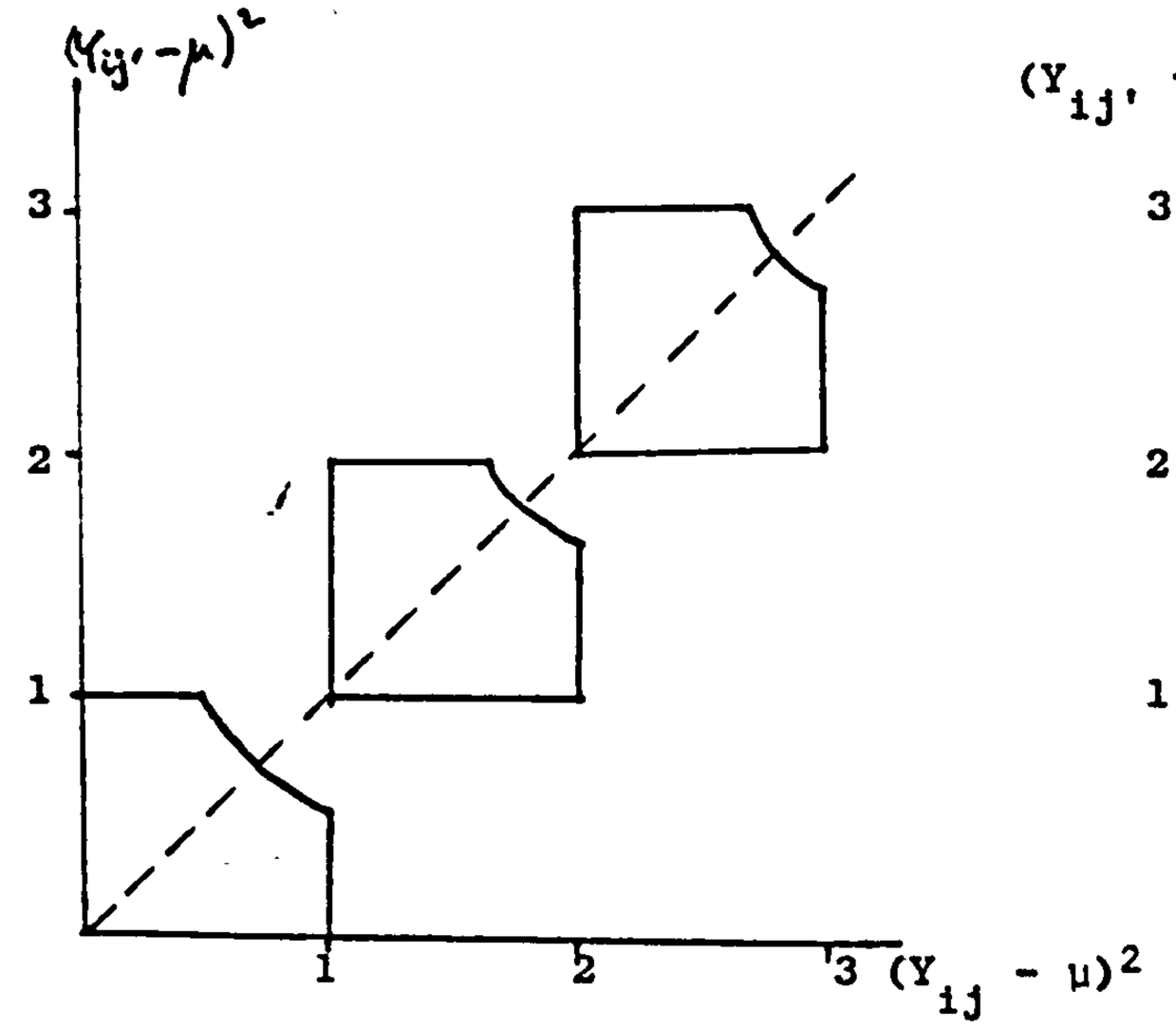
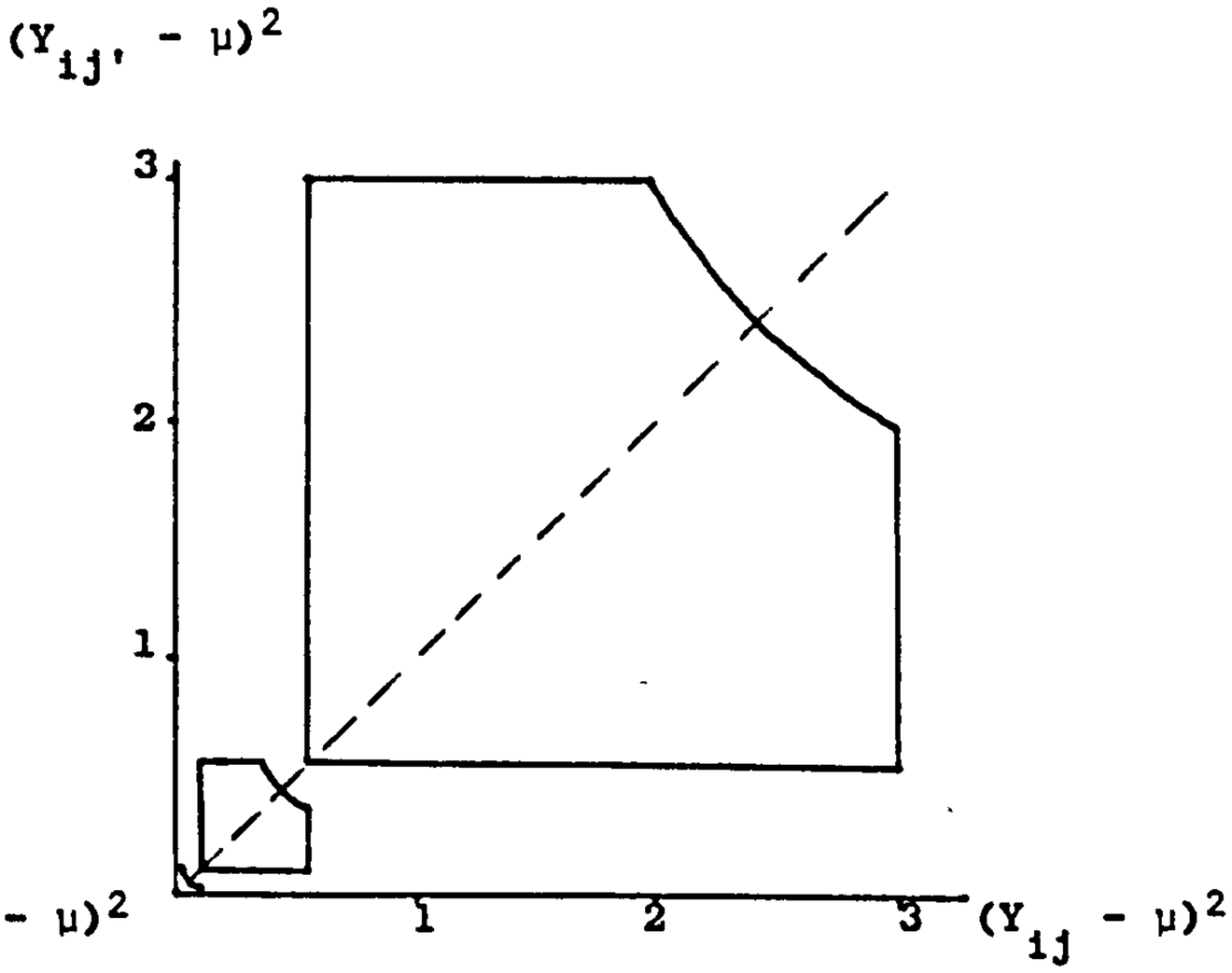


Figure 5.32 - Population 2
95% probability regions for
$(Y - \mu)^2$, $(Y' - \mu)^2$



$\tau_{Ym} < {}^{1}/\sqrt{3} \cdot \tau_{Y\tilde{v}}$ increases as c_1 or r_v increases.

Figures 5.29 - 5.32 have been drawn with high values of τ_{Ym} for graphical simplicity and the latter case applies. However, in practice τ_{Ym} is unlikely to be greater than .58. This example illustrates the dangers of drawing general conclusions from such graphical comparisons.

Case 4 : B holds but A does not hold

There seems to be little point in extending the algebraic or graphical comparisons of the last sections, since in this case we would need to model several functions of M, for example $\sigma_B^2(M)$, $c_1(M)$ and $\gamma(M)$, even when normality is assumed. However, it is possible to extend the spatial process argument of Section 5.3

We may generalise (5.39) to

$$\tau_{hi} = \int_0^1 \rho_h(\alpha_i\ t)\ K(t)\ dt \qquad (5.56)$$

where $\rho_h(s) = \text{corr}\left[h(Y(\underline{x})),\ h(Y(\underline{x}'))\right]$

and s is the distance between $\underline{x}$ and $\underline{x}'$.

For a stationary stochastic process

$$\rho(s) = \frac{k_{11}}{k_2} \qquad (5.57)$$

$$\rho_{hv}(s) = \frac{\bar{k}_{22} + 2k_{11}^2}{\bar{k}_4 + 2\bar{k}_2}$$

where k_2, $\bar{k}_4$ are the (common) univariate cumulants of $Y(\underline{x})$ and $Y(\underline{x}')$ and k_{11} and $\bar{k}_{22}$ are bivariate cumulants. In the case of a Gaussian process $k_{22} = k_4 = 0$ and

$$\rho_{hv}(s) = \rho(s)^2 \quad \text{uniformly in } s$$

Hence

$$\tau_{Yvi} = \int_0^1 \rho^2(\alpha_i t)\, K(t)\, dt \qquad (5.58)$$

$$\lesssim \tau_{Ymi} \quad \text{from (5.39)}$$

Hence Kish and Frankel's (1974) conjecture that deffs of complex statistics are less than deffs of means applies. In the non-gaussian case we may write $\rho_{hv}(s)$ in terms of $\rho(s)$ as in (5.52) and argue as in Case 1 that $\rho_{hv}(s)$ will only be greater than $\rho(s)$ in very extreme cases, so that again we would expect the conjecture to apply. We consider two examples.

Example 1: $$\rho(s) = as^{-b} \qquad (5.59)$$

$$\rho_{hv}(s) = a^2 s^{-2b}$$

$$\alpha_i \propto M_i^{\frac{1}{2}}$$

Then $$\tau_{Ym}(M_i) \propto M_i^{-b/2}$$

$$\tau_{Y\tilde{v}}(M_i) \propto M_i^{-b}$$

The constants of proportionality may be taken as unity if we assume $\tau_{Ym}(1) = \tau_{Y\tilde{v}}(1) = 1$. Hence not only is $\tau_{Y\tilde{v}i}$ less than τ_{Ymi} uniformly in M_i but $\tau_{Y\tilde{v}i}$ also decreases at a faster rate than τ_{Ymi} as M_i increases.

Note,however, that the analogy with the model of Smith (1938) described in Section 5.3 seems to fail. From (5.34)

$$\tau_{Ym}(M_i) = V_I(\mu_i|M_i)/\sigma^2$$

$$\doteq V_I(\bar{Y}_i|M_i)/\sigma^2$$

Smith (1938) argued reasonably that $V_I(\bar{Y}_i|M_i) \geq \sigma^2/M_i$ so that the coefficient of M_i in $V_I(\bar{Y}_i|M_i)$ should lie between -1 and 0. In our notation $0 \leq b \leq 2$. But the same argument would suggest that the coefficient of M_i in $\tau_{Y\tilde{v}}(M_i)$ should lie between -1 and 0 i.e. $0 \leq b \leq 1$ which is inconsistent.

Example 2

$$\rho_{h1}(s) = e^{-\lambda s} \tag{5.60}$$

$$\rho_{h2}(s) = e^{-2\lambda s}$$

$$\alpha_i \propto M_i^{\frac{1}{2}}$$

$$K(t) \sim \text{gamma } (\beta, r)$$

Then

$$\tau_{Ymi} \propto (\lambda M_i^{\frac{1}{2}} + \beta)^{-r}$$

$$\tau_{Y\tilde{v}i} \propto (2\lambda M_i^{\frac{1}{2}} + \beta)^{-r}$$

Assuming that the effect of β is negligible for large M_i

$$\tau_{Ymi} \propto M_i^{-r/2}$$

$$\tau_{Y\tilde{v}i} \propto M_i^{-r/2}$$

In this example $\tau_{Y\tilde{v}i}$ is again uniformly less than τ_{Ymi} but now decreases at the same rate.

We now establish some preliminary results in order to prove Theorem 5.23, which shows that the meffs of T_{Yv} and $T_{Y\tilde{v}}$ are asymptotically equivalent in a certain sense. First of all Lemma 5.19 gives formulae for the first two moments of a quadratic form. These formulae reduce to the standard results for normal random variables (e.g. Searle, 1971, p.57) as a special case.

Lemma 5.19

Let $X_1 \ldots X_n$ be independent random variables with zero means, variances σ_i^2 and fourth cumulants, k_{4i} (i=1 ...n).

Let A_{ij} (i,j=1 ... n) be constants.

Then $E(\sum_{ij} A_{ij} X_i X_j) = \sum_i A_{ii} \sigma_i^2$

$$\text{var}(\sum_{ij} A_{ij} X_i X_j) = 2\sum_{ij} A_{ij}^2 \sigma_i^2 \sigma_j^2 + \sum_i A_{ii}^2 k_{4i}$$

Corollary 5.20

If $A_{ij} = w_i w_j$ and $Y = \Sigma w_i X_i$ then

$$E(Y^2) = \Sigma w_i^2 \sigma_i^2$$

$$\text{var}\,(Y^2) = 2(\Sigma w_i^2 \sigma_i^2)^2 + \Sigma w_i^4 k_{4i}$$

In order to prove Lemma 5.6, we shall require the following generalisation of Cauchy's inequality.

Lemma 5.21

Let $Z_1 \ldots Z_n$ be random variables (not necessarily independent) with finite n^{th} moments.

Then

$$E\left[\prod_{i=1}^{n} Z_i\right] \leq \left[\prod_{i=1}^{n} E(Z_i^n)\right]^{1/n} \qquad (5.61)$$

Proof

(5.61) is true for n=2 since it is then equivalent to Cauchy's inequality.

Suppose (5.61) is true for n.

Then

$$E\left[\prod_{i=1}^{n+1} Z_i\right] \leq E\left[\left(\prod_{i=1}^{n} X_i\right)^{\frac{n+1}{n}}\right]^{\frac{n}{n+1}} \left[E(X_{n+1}^{n+1})\right]^{\frac{1}{n+1}}$$

$$\leq \left\{\prod_{i=1}^{n} E\left[(X_i^{\frac{n+1}{n}})^n\right]\right\}^{\frac{1}{n+1}} \left\{E(X_{n+1}^{n+1})\right\}^{\frac{1}{n+1}} \quad \text{by Holder's inequality}$$

$$= \left[\prod_{i=1}^{n+1} E(X_i^{n+1})\right]^{\frac{1}{n+1}} \quad \text{by assumption}$$

Hence (5.61) holds for n+1 and Lemma 5.21 follows by Induction.

Lemma 5.22

Let $Z_1 \ldots Z_m$ be random variables (not necessarily independent) which have a common marginal distribution with moments:

$$\mu_r = E(Z_i^r) < \infty \quad r=1,2 \ldots$$

and cumulants k_r; $r=1,2 \ldots$

Let $\overline{Z} = \Sigma Z_i/m$ have moments $\overline{\mu}_r$ and cumulants $\overline{k}_r$.

Then

$$\overline{\mu}_r \leq \mu_r$$

$$\overline{k}_r \leq k_r \quad r=1,2 \ldots$$

Proof

Let $\phi_Z(t)$ be the characteristic function of the common marginal distribution of $Z_1 \ldots Z_m$ and let $\psi_Z(t)$ be the corresponding cumulant generating function.

$$\phi_Z(t) = E(e^{iZ_j t})$$

$$\psi_Z(t) = \log \phi_Z(t)$$

Similarly let

$$\phi_{\overline{Z}}(t) = E(e^{i\overline{Z}t})$$

$$\psi_{\overline{Z}}(t) = \log \phi_{\overline{Z}}(t)$$

Then

$$\phi_{\overline{Z}}(t) = E(\prod_j e^{iZ_j t/m})$$

$$\leq \left[\prod_j E(e^{iZ_j t})\right]^{1/m} \text{ by Lemma 5.5}$$

$$= \prod_j \phi_Z(t)^{1/m}$$

$$= \phi_Z(t) \text{ uniformly in } t$$

Hence $\overline{\mu}_r \leq \mu_r \quad r=1,2 \ldots$

and since log is a monotonic increasing function,

$$\psi_{\overline{Z}}(t) = \log \phi_{\overline{Z}}(t)$$

$$\leq \log \phi_Z(t)$$

$$= \psi_Z(t)$$

and $\overline{k}_r \leq k_r \quad r=1,2 \ldots$

In order to demonstrate the asymptotic equivalence of $\text{meff}(T_{Yv})$ and $\text{meff}(T_{Yv})$ we recall the limiting argument of Section 5.2. We consider a nested sequence of finite populations U_n obeying Model I and a sequence of designs $p_n(s|\underline{M})$ selecting a fixed number, n , of clusters.

We assume

C4 : $\sum_{i\varepsilon s} m_i^r/n \xrightarrow{a.s} \mu_{mr} < \infty$ as $N \to \infty$, $r = 1, 2, 4$. where the limit is taken with respect to $p_n(s|\underline{M})$ and $h_c(M_i)$. Assumption C4 seems reasonable since $m_i \leq M_i$ and so

$$\Sigma m_i^r/n \leq \Sigma M_i^r \quad /n$$

and by the Strong Law of Large Numbers $\Sigma M_i^r/n$ converges almost surely provided appropriate moments of $h_c(M)$ are finite.

Hence $\Sigma m_i^r/n$ is bounded above and below and so will only fail to converge if the sampling designs depend on N in a peculiar 'non-monotonic' manner.

Theorem 5.23

If Assumptions B and C4 hold then

$$\text{meff}(T_{Yv}|s,\underline{M}) - \text{meff}(T_{Yv}|s,\underline{M}) \xrightarrow{a.s.} 0 \quad \text{as} \quad N \to \infty$$

Proof Let $w_i = m_i/m_o$

$$Z_{ij} = Y_{ij} - \mu$$

$$X_i = \sum_{j=1}^{m_i} Z_{ij}/m_i$$

Then

$$T_{Ym} - \mu = \sum_{i=1}^{n} w_i X_i$$

and

$$T_{Yv} = \Sigma\Sigma(Y_{ij} - T_{Ym})^2/(m_o - 1)$$

$$= T_{Y\tilde{v}} - m_o(T_{Ym} - \mu)^2/(m_o - 1)$$

$$= T_{Y\tilde{v}} - m_o(\Sigma w_i X_i)^2/(m_o - 1) \qquad (5.62)$$

Under Model I, $Z_{i1} \ldots Z_{imi}$ have a common marginal distribution (given s and $\underline{M}$), if B holds, with mean zero, variance σ^2 and fourth cumulant k_4 for i=1 ... n. Let σ^2_{Xi} and k_{4Xi} be the variance and fourth cumulant of X_i (given s and $\underline{M}$). Then from Lemma 5.22

$$\sigma_{Xi}^2 \leq \sigma^2$$

$$k_{4Xi} \leq k_4$$

Now $X_1 \ldots X_n$ are independent (given s and $\underline{M}$), with zero means.

Hence from corollary 5.20

$$\text{var}_I\left[(\Sigma w_i X_i)^2 \middle| s, \underline{M}\right] = 2(\Sigma w_i^2 \sigma_{Xi}^2)^2 + \Sigma w_i^4 k_{4Xi}$$

$$\leq 2\sigma^4(\Sigma w_i^2)^2 + (\Sigma w_i^4)\, k_4$$

$$= 2\sigma^4(\Sigma m_i^2)^2/m_o^4 + (\Sigma m_i^4)\, k_4/m_o^4$$

It follows from Assumption C4 that

$$n^2\text{var}_I\left[(\Sigma w_i X_i)^2 \middle| s, \underline{M}\right] \xrightarrow{a.s.} C_1 = 2\sigma^4 \mu_{m2}^2/\mu_{m1}^4 \qquad (5.63)$$

where C_1 is a constant

From Lemma 5.2 we obtain

$$n\text{var}_I\left[T_{Y\tilde{v}} \middle| s, \underline{M}\right] = nC_2(\Sigma m_i + \Sigma m_i(m_i-1)\,\tau_{Y\tilde{v}})/(\Sigma m_i - 1)^2$$

$$\xrightarrow{a.s.} C_2(\mu_{m1} + (\mu_{m2} - \mu_{m1})\tau_{Y\tilde{v}})/\mu_{m1}^2$$

$$= C_3 \qquad (5.64)$$

where C_2 and C_3 are constants.

From (5.62) we obtain

$$n\left[\text{var}_I\,(T_{Y\tilde{v}} \middle| s, \underline{M}) - \text{var}_I\,(T_{Y\tilde{v}} \middle| s, \underline{M})\right]$$

$$= n\left[\frac{m_o^2}{(m_o-1)^2}\ \text{var}_I\left[(\Sigma w_i X_i)^2 \middle| s, \underline{M}\right]\right.$$

$$\left. - \frac{2m_o}{m_o-1}\ \text{cov}_I(T_{Y\tilde{v}},\ (\Sigma w_i X_i)^2 \middle| s, \underline{M})\right]$$

$$\leq \frac{m_o^2 n}{(m_o-1)^2}\ \text{var}_I\ \left[(\Sigma w_i X_i)^2 \middle| s, \underline{M}\right]$$

$$+ \frac{2m_o}{m_o-1}\ \left[\text{var}_I(T_{Y\tilde{v}} \middle| s, \underline{M})\ \text{var}_I((\Sigma w_i X_i)^2 \middle| s, \underline{M})\right]^{\frac{1}{2}}$$

$$\xrightarrow{a.s.} 0 \qquad (5.65)$$

from (5.63) and (5.64).

But the previous argument is also valid for Model II. Hence

$$\text{nvar}_{II}[T_{Y\tilde{v}}|\, s,\, \underline{M}\,] \xrightarrow{a.s.} C_4 \quad , \text{ a constant}$$

and

$$n[\text{var}_{II}(T_{Yv}|\, s,\, \underline{M}\,) - \text{var}_{II}(T_{Y\tilde{v}}|\, s,\, \underline{M}\,)] \xrightarrow{a.s.} 0 \qquad (5.66)$$

The result then follows from (5.65) and (5.66) using Definition 5.1 and noting that C_3 and C_4 are non-zero.

Theorem 5.23 demonstrates the almost-sure convergence of the meffs of T_{Yv} and $T_{Y\tilde{v}}$ as n increases. The rate of convergence is of order $O_p(n^{-\frac{1}{2}})$

Finally, to make a finite sample comparison between the two meffs we give the exact meff of T_{Yv} in Theorem 5.24.

Theorem 5.24

If B holds

$$E_I(T_{Yv}|s,\underline{M},\psi) = \sigma^2(1 - \sum_{1}^{n} m_i(m_i-1)\tau_{Ym}(M_i)/(m_o-1)m_o) \qquad (5.67)$$

$$E_{II}(T_{Yv}|s,\underline{M},\psi) = \sigma^2 \qquad (5.68)$$

and the (conditional) meff of T_{Yv} is

$$\text{meff}(T_{Yv}|s,\underline{M},\psi) = \text{var}_I(T_{Yv}|s,\underline{M})/\text{var}_{II}(T_{Yv}|^s,\underline{M}) \qquad (5.69)$$

where

$$\text{var}_I(T_{Yv}|s,\underline{M}) = \Big[\sum_{i=1}^{n} (2m_o m_i^2(m_o - 2m_i)\sigma_B^4(M_i)$$

$$+ 4\, m_o m_i(m_o - 2m_i)\, \sigma_B^2(M_i)\sigma_W^2(M_i)$$

$$+ 2\, m_o m_i(m_o - 2)\, \sigma_W^4(M_i)$$

$$+ m_i^2(m_o - m_i)^2 k_{4B}(M_i) + m_i(m_o - 1)^2 k_{4W}(M_i)$$

$$+ m_i(m_o^2 m_i + 2m_o^2 - 2m_o m_i - 4m_o + 3m_i)\gamma(M_i)$$

$$+ m_i(4m_o^2 + 2m_o^2 m_i - 2m_o m_i^2 - 10m_o m_i + 6m_i^2)c_1(M_i)$$

$$+ 4m_i(m_o - 1)(m_o - m_i)\, c_2(M_i))$$

$$+ (m_o + \Sigma m_i(m_i - 1)\tau_{Ymi})^2\Big]/m_o(m_o - 1)^2 \qquad (5.70)$$

$$\text{var}_{II}(T_{Yv}|s,\underline{M}) = k_4/m_o + 2\sigma^4/(m_o-1) \qquad (5.71)$$

Proof: Using the same notation as in Theorem 5.23 we first establish (5.67)

$$E_I[(\Sigma w_i X_i)^2|s,\underline{M}] = \Sigma w_i^2 \sigma_{Xi}^2 \quad \text{from Corollary 5.20}$$

where $\sigma_{Xi}^2 = E_I(X_i^2|M_i)$

$$= E_I[E((X_i-\mu_i)^2 + (\mu_i-\mu)^2|\theta_i) \mid M_i]$$

$$= E_I[\sigma_i^2/m_i + (\mu_i-\mu)^2|M_i]$$

$$= \sigma_W^2(M_i)/m_i + \sigma_B^2(M_i)$$

$$= \sigma^2/m_i + (m_i-1)\ \sigma_B^2(M_i)/m_i$$

$$= \sigma^2(1 + (m_i-1)\ \tau_{Ym}(M_i))/m_i \quad \text{from (5.31)} \qquad (5.72)$$

Hence

$$E_I\left[(\Sigma w_i X_i)^2 \middle| s,\underline{M}\right] = \sigma^2\ \Sigma m_i(1+(m_i-1)\tau_{Ym}(M_i))/m_o^2$$

$$= \sigma^2(m_o+\Sigma m_i(m_i-1)\tau_{Ym}(M_i))/m_o^2 \qquad (5.73)$$

Hence from (5.40a) and (5.62)

$$E_I\left[T_{Yv} \middle| s,\underline{M}\right] = m_o\sigma^2/(m_o-1) - \sigma^2(m_o + \Sigma m_i(m_i-1)\tau_{Ym}(M_i))/m_o(m_o-1)$$

$$= \sigma^2(1-\Sigma m_i(m_i-1)\tau_{Ym.}(M_i)/m_o(m_o-1))$$

We now evaluate the variance of T_{Yv} under Model I. From (5.62) we have

$$\mathrm{var}_I\left[T_{Yv} \middle| s,\underline{M}\right] = \mathrm{var}_I\left[T_{Y\tilde{v}} \middle| s,\underline{M}\right] - 2m_o\ \mathrm{cov}\left[T_{Y\tilde{v}},(\Sigma w_i X_i)^2 \middle| s,\underline{M}\right]/(m_o-1)$$

$$+ m_o^2\ \mathrm{var}_I\left[(\Sigma w_i X_i)^2 \middle| s,\underline{M}\right]/(m_o-1)^2 \qquad (5.74)$$

And from Corollary 5.20

$$\mathrm{var}_I\left[(\Sigma w_i X_i)^2 \middle| s,\underline{M}\right] = 2(\Sigma w_i^2\sigma_{Xi}^2)^2 + \Sigma w_i^4 k_{4Xi} \qquad (5.75)$$

where $k_{4Xi} = E_I\left[X_i^4 \middle| M_i\right] - 3\sigma_{Xi}^4$

$$= E_I\left[E\left[(X_i-\mu_i)^4 + 4(X_i-\mu_i)^3(\mu_i-\mu) + 6(X_i-\mu_i)^2(\mu_i-\mu)^2\right.\right.$$

$$\left.\left. + 4(X_i-\mu_i)(\mu_i-\mu)^3 + (\mu_i-\mu)^4 \middle| \theta_i\right] \middle| M_i\right] - 3\sigma_{Xi}^4$$

$$= E_I\left[3\sigma_i^4/m_i^2 + k_{4i}/m_i^3 + 4k_{3i}(\mu_i-\mu)/m_i^2 + 6\sigma_i^2(\mu_i-\mu)^2/m_i\right.$$

$$\left. + (\mu_i-\mu)^4 \middle| M_i\right] - 3\sigma_{Xi}^4$$

$$= (3\sigma_W^4(M_i) + 3\gamma(M_i))/m_i^2 + k_{4W}(M_i)/m_i^3 + 4c_2(M_i)/m_i^2$$

$$+ 6(c_1(M_i) + \sigma_W^2(M_i)\sigma_B^2(M_i))/m_i + k_{4B}(M_i) + 3\sigma_B^4(M_i)$$

$$- 3(\sigma_W^2(M_i)/m_i + \sigma_B^2(M_i))^2$$

$$= k_{4B}(M_i) + 6c_1(M_i)/m_i + 3\gamma(M_i)/m_i^2 + 4c_2(M_i)/m_i^2$$

$$+ k_{4W}(M_i)/m_i^3 \qquad (5.76)$$

We note, as a check, that $k_{4Xi} = k_4$ if $m_i = 1$ from Lemma 5.16.

Combining (5.73), (5.75) and (5.76), we obtain

$$\text{var}_I\left[(\Sigma w_i X_i)^2 | s, \underline{M}\right] = 2\sigma^4(m_o + \Sigma m_i(m_i-1)\tau_{Ym}(M_i))^2/m_o^4$$

$$+ \Sigma m_i^4(k_{4B}(M_i) + 6c_1(M_i)/m_i + 3\gamma(M_i)/m_i^2$$

$$+ 4c_2(M_i)/m_i^2 + k_{4W}(M_i)/m_i^3)/m_o^4 \qquad (5.77)$$

We now evaluate the second term in (5.74)

$$\text{cov}_I\left[T_{Y\tilde{v}}, (\Sigma w_i X_i)^2 | s, \underline{M}\right] = \text{cov}_I\left[\Sigma\Sigma Z_{ij}^2, (\Sigma w_i X_i)^2 | s, \underline{M}\right]/(m_o-1)$$

$$= \Sigma\Sigma w_i^2 \, \text{cov}_I(Z_{ij}^2, X_i^2 | s, \underline{M})/(m_o-1)$$

(5.78)

since $\text{cov}_I(Z_{ij}^2, w_i X_i w_k X_k | s, \underline{M}) = 0$ if $i \neq k$

Now

$$\text{cov}_I(Z_{ij}^2, X_i^2 | s, \underline{M}) = \text{cov}_I\left[E_I(Z_{ij}^2 | \theta_i), E(X_i^2 | \theta_i) | s, \underline{M}\right]$$

$$+ E_I\left[\text{cov}(Z_{ij}^2, X_i^2 | \theta_i) | s, \underline{M}\right] \qquad (5.79)$$

We shall require the following moments of Z_{ij}

$$E_I(Z_{ij} | \theta_i) = \mu_i - \mu \qquad (5.80)$$

$$E_I(Z_{ij}^2 | \theta_i) = (\mu_i-\mu)^2 + \sigma_i^2 \qquad (5.81)$$

$$E_I(Z_{ij}^3 | \theta_i) = (\mu_i-\mu)^3 + 3(\mu_i-\mu)\sigma_i^2 + k_{3i} \qquad (5.82)$$

$$E_I(Z_{ij}^4 | \theta_i) = (\mu_i-\mu)^4 + 6(\mu_i-\mu)^2\sigma_i^2 + 4(\mu_i-\mu)k_{3i} + k_{4i} + 3\sigma_i^4 \qquad (5.83)$$

Furthermore

$$E_I(X_i^2|\theta_i) = (\mu_i - \mu)^2 + \sigma_i^2/m_i$$

and so from (5.81)

$$\text{cov}_I\left[E_I(Z_{ij}^2|\theta_i)\ ,\ E(X_i^2|\theta_i)|s,\underline{M}\right] = \text{cov}_I\left[(\mu_i-\mu)^2 + \sigma_i^2,\ (\mu_i-\mu)^2 + \sigma_i^2/m_i|s,\underline{M}\right]$$

$$= k_{4B}(M_i) + 2\sigma_B^4(M_i) + (m_i + 1)\ c_1(M_i)/m_i + \gamma(M_i)/m_i \qquad (5.84)$$

Now

$$\text{cov}_I(Z_{ij}^2,\ X_i^2|\theta_i) = \text{cov}_I(Z_{ij}^2, \sum_k Z_{ik}^2 + \sum_{k\neq k'} Z_{ik}Z_{ik'}|\theta_i)/m_i^2$$

$$= \left[\text{var}_I(Z_{ij}^2|\theta_i) + 2(m_i-1)\ \text{cov}_I(Z_{ij}^2,\ Z_{ij}|\theta_i)\ E_I(Z_{ij}|\theta_i)\right]/m_i^2$$

$$= \left[4(\mu_i - \mu)^2\sigma_i^2 + 4(\mu_i - \mu)\ k_{3i} + k_{4i} + 2\sigma_i^4 + 2(m_i-1)(2(\mu_i - \mu)\sigma_i^2 + k_{3i})(\mu_i - \mu)\right]/m_i^2$$

from (5.78) - (5.81)

$$= \left[4m_i(\mu_i - \mu)^2\sigma_i^2 + 2(m_i + 1)(\mu_i - \mu)\ k_{3i} + k_{4i} + 2\sigma_i^4\right]/m_i^2$$

Hence

$$E_I\left[\text{cov}_I(Z_{ij}^2\ ,\ X_i^2|\theta_i)|s,\underline{M}\right] = \left[4m_i(\sigma_B^2(M_i)\sigma_W^2(M_i) + c_1(M_i)) + 2(m_i + 1)\ c_2(M_i) + k_{4W}(M_i) + 2\gamma(M_i) + 2\sigma_W^4(M_i)\right]/m_i^2 \qquad (5.85)$$

Substituting (5.84) and (5.75) into (5.79)

$$\text{cov}_I(Z_{ij}^2,\ X_i^2|s,\underline{M}) = k_{4B}(M_i) + 2(\sigma_B^2(M_i) + \sigma_W^2(M_i)/m_i)^2 + (m_i + 5)\ c_1(M_1)/m_i + (m_i + 2)\gamma(M_i)/m_i^2 + 2(m_i + 1)c_2(M_i)/m_i^2 + k_{4W}(M_i)/m_i^2 \qquad (5.86)$$

Substituting (5.86) into (5.78)

$$\text{cov}_I\left[T_{Y\tilde{v}},\ (\Sigma w_i X_i)^2 \mid s, \underline{M}\right] = \sum_i (m_i^3 k_{4B}(M_i) + 2m_i^3(\sigma_B^2(M_i) + \sigma_W^2(M_i)/m_i)^2$$

$$+ (m_i + 5)\, m_i^2 c_1(M_i) + m_i(m_i + 2)\gamma(M_i) + 2m_i(m_i + 1)\, c_2(M_i)$$

$$+ m_i k_{4W}(M_i))/m_o^2\ (m_o - 1) \qquad (5.87)$$

Now

$$\text{var}_{II}(T_{Y\tilde{v}} \mid s, \underline{M}) = m_o\ \text{var}\ (Z_{ij}^2 \mid M_i)/(m_o - 1)^2$$

$$= m_o(k_4 + 2\sigma^4)/(m_o - 1)^2 \qquad (5.88)$$

Hence from (5.41)

$$\text{var}_I(T_{Y\tilde{v}} \mid s, \underline{M}) = (1 + \Sigma m_i(m_i - 1)\tau_{Y\tilde{v}}(M_i)/m_o)m_o(k_4 + 2\sigma^4)/(m_o - 1)^2$$

$$= [m_o(k_4 + 2\sigma^4) + \Sigma m_i(m_i - 1)(2\sigma_B^4(M_i) +$$

$$k_{4B}(M_i) + 2c_1(M_i) + \gamma(M_i))]/(m_o - 1)^2 \qquad (5.89)$$

from Lemma 5.17

Substituting (5.77), (5.87) and (5.89) into (5.74) we obtain (5.70). The momentsof T_{Yv} under Model II in (5.68) and (5.71) are obtained directly from classical theory.

In Table 5.3 we give estimates of the misspecification effects of $T_{Yv}^{\sim}$ and T_{Yv} for the Family Expenditure Survey data. The estimation procedure is described in the Appendix. It is clear that for this data the estimates are very close.

Table 5.3. Estimates of misspecification effects for FES data

Variable Y	meff($T_{Yv}^{\sim}$)	meff(T_{Yv})
log (V1)	1.1366	1.1374
log (V2)	2.1123	2.1121
log (V3)	1.3004	1.3005

5.5 Misspecification Effects of Covariances

It is now sufficient to assume that y_{ij} is bivariate, i.e. $p = 2$. We label the components (x_{ij}, y_{ij}). Let

$$T_{XYc} = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij}-T_{Xm})(y_{ij}-T_{Ym})/(m_o-1)$$

where

$$T_{Xm} = \Sigma\Sigma x_{ij}/m_o \quad , \quad T_{Ym} = \Sigma\Sigma y_{ij}/m_o$$

We consider T_{XYc} as an estimator (the standard estimator) of σ_{XY}, the covariance between X and Y for f_o defined in (5.3). T_{XYc} is, of course, a generalisation of T_{Yv} defined in (5.40) since $T_{Yv}=T_{YYc}$. As for T_{Yv} we shall approximate T_{XYc} by

$$T_{XY\tilde{c}} = \sum_{i=1}^{n} \sum_{j=1}^{\dot{m}_i} h_c(x_{ij}, y_{ij})$$

where

$$h_c(x_{ij},y_{ij}) = (x_{ij}-\mu_X)(y_{ij}-\mu_Y)/(m_o-1)$$

and μ_X and μ_Y are the means of X and Y respectively in f_o.

We shall approximate the misspecification effect of T_{XYc} by that of $T_{XY\tilde{c}}$ which may be obtained using the results of Section 5.2 since $T_{XY\tilde{c}}$ is of the additive form (5.6). From Lemma 5.2 it follows that the effect of misspecifying Model I as Model II does not introduce any bias into $T_{XY\tilde{c}}$ under Assumption B.

$$E_I(T_{XY\tilde{c}}|s,\underline{M}) = E_{II}(T_{XY\tilde{c}}|s,\underline{M}) = m_o\sigma_{XY}/(m_o-1)$$

Lemma 5.25

If Assumption B holds

$$\text{meff}(T_{XY\tilde{c}}|s,\underline{M}) = 1 + \Sigma\, m_i(m_i-1)\,\tau_{XY\tilde{c}}(M_i)/m_o \qquad (5.90)$$

where

$$\tau_{XY\tilde{c}}(M_i) = \text{corr}_I\left[h_c(X_{ij},Y_{ij}),\ h_c(X_{ij'},Y_{ij'})\,|M_i\right] \quad j \neq j' \qquad (5.91)$$

Corollary 5.26

If Assumption A holds

$$\text{meff}(T_{XY\tilde{c}}|s,\underline{M}) = 1 + (m^*-1)\tau_{XY\tilde{c}} \qquad (5.92)$$

In order to express $\tau_{XY\tilde{c}}$ in terms of θ_i we introduce the following notation.

Let $\mu_{Xi},\mu_{Yi},\sigma^2_{Xi},\sigma^2_{Yi}$ be the univariate within-cluster moments as defined in Section 5.4.

Let

$$\sigma_{XYi} = \text{cov}_I(X_{ij},Y_{ij}|\theta_i)$$

Let $\sigma^2_{XB}(M_i)$, $\sigma^2_{YB}(M_i)$, $\sigma^2_{XW}(M_i)$, $\sigma^2_{YW}(M_i)$ be defined as in Section 5.4.

Let

$$\sigma_{XYB}(M_i) = \text{cov}_I(\mu_{Xi},\mu_{Yi}|M_i)$$

$$\sigma_{XYW}(M_i) = E_I(\sigma_{XYi}|M_i)$$

$$\gamma_{XY}(M_i) = \text{var}_I(\sigma_{XYi}|M_i)$$

$$c_{1XY}(M_i) = \text{cov}_I(\sigma_{XYi},\ (\mu_{Xi}-\mu_X)(\mu_{Yi}-\mu_Y)|M_i)$$

$$k_{22B}(M_i) = E_I[(\mu_{Xi}-\mu_X)^2(\mu_{Yi}-\mu_Y)^2|M_i]$$

$$- \sigma^2_{XB}(M_i)\ \sigma^2_{YB}(M_i) - 2\sigma^2_{XYB}(M_i)$$

k_{22B} is the $(2,2)^{th}$ bivariate cumulant of $(\mu_{Xi},\ \mu_{Yi})$ (Kendall and Stuart, 1969, p.82) and is equal to zero if μ_{Xi} and μ_{Yi} are jointly normally distributed.

It follows as in Section 5.4 that if B holds then

$$\sigma_X^2 = \sigma^2_{XB}(M_i) + \sigma^2_{XW}(M_i)$$

$$\sigma_Y^2 = \sigma^2_{YB}(M_i) + \sigma^2_{YW}(M_i)$$

$$\sigma_{XY} = \sigma_{XYB}(M_i) + \sigma_{XYW}(M_i)$$

If B holds σ_X^2, σ_Y^2 and σ_{XY} do not depend on M_i, as does not the $(2,2)^{th}$ cumulant of (X_{ij}, Y_{ij})

$$k_{22} = E_I\left[(X_{ij}-\mu_X)^2(Y_{ij}-\mu_Y)^2 \middle| M_i\right] - \sigma_X{}^2\sigma_Y{}^2 - 2\sigma_{XY}{}^2$$

Lemma 5.27

If B holds

$$\tau_{XY\tilde{c}}(M_i) = \frac{\sigma_{XB}^2(M_i)\sigma_{YB}^2(M_i) + \sigma_{XYB}^2(M_i) + k_{22B}(M_i) + 2c_{1XY}(M_i) + \gamma_{XY}(M_i)}{\sigma_X{}^2\,\sigma_Y{}^2 + \sigma_{XY}{}^2 + k_{22}} \qquad (5.93)$$

Proof:

From (5.91)

$$\tau_{XY\tilde{c}i} = \mathrm{corr}_I\left[(X_{ij}-\mu_X)(Y_{ij}-\mu_Y),(X_{ij'}-\mu_X)(Y_{ij'}-\mu_Y) \middle| M_i\right] \quad j \neq j'$$

$$= \frac{\mathrm{var}_I\left[(\mu_{Xi}-\mu_X)(\mu_{Yi}-\mu_Y) + \sigma_{XYi} \middle| M_i\right]}{\mathrm{var}_I\left[(X_{ij}-\mu_X)(Y_{ij}-\mu_Y) \middle| M_i\right]}$$

$$= \frac{\sigma_{XB}{}^2(M_i)\sigma_{YB}^2(M_i) + \sigma_{XYB}^2(M_i) + k_{22B}(M_i) + 2c_{1XY}(M_i) + \gamma_{XY}(M_i)}{\sigma_X{}^2\,\sigma_Y{}^2 + \sigma_{XY}{}^2 + k_{22}}$$

Corollary 5.28

If A holds

$$\tau_{XYc}^{\sim} = \frac{\sigma_{XB}^2 \sigma_{YB}^2 + \sigma_{XYB}^2 + k_{22B} + 2c_{1XY} + \gamma_{XY}}{\sigma_X^2 \sigma_Y^2 + \sigma_{XY}^2 + k_{22}} \qquad (5.94)$$

We note that, in the special case when $X_{ij} = Y_{ij}$,

$$\sigma_{XB}^2 = \sigma_{YB}^2 = \sigma_B^2 \qquad \sigma_{XYB}^2 = \sigma_B^4$$

$$\sigma_X^2 = \sigma_Y^2 = \sigma^2 \qquad \sigma_{XY}^2 = \sigma^4$$

$$k_{22B} = k_{4B} \qquad k_{22} = k_4$$

$$c_{1XY} = c_1 \qquad \gamma_{XY} = \gamma$$

$$h_c(X,Y) = h_v(X)$$

and Lemmas 5.25 and 5.27 reduce to Lemmas 5.14 and 5.17.

We now consider diagnostic checks of Assumptions B and A. Univariate checks were given in Section 5.3 and 5.4. In order for there to be no misspecification bias in $T_{XYc}^{\sim}$ we need $E_I\left(h_c(X_{ij},Y_{ij})|M_i\right)$ to be free of M_i. This requirement may be checked by plotting

$$(m_o-1)\hat{\bar{h}}_{ci} = (m_o-1)\sum_{j=1}^{m_i} \hat{h}_c(x_{ij},y_{ij})/m_o$$

$$= \sum_{j=1}^{m_i} (x_{ij}-\hat{\mu}_X)(y_{ij}-\hat{\mu}_Y)/m_o ,$$

where $\hat{\mu}_X$ and $\hat{\mu}_Y$ are defined as in (5.47), against M_i. If B holds the regression function $E_I\left((m_o-1)\hat{\bar{h}}_{ci}|M_i\right)$ should not depend on M_i (assuming the effect of estimating μ_X and μ_Y is negligible). Such plots are given in Figures 5.33 - 5.35 for the National Survey of Attainment data. Although the variance functions do appear to depend on M_i, there is very little visual evidence of the regression functions depending on M_i. We note that the negative covariances in Figure 5.33 are accounted for by the fact that AM-Attitude to Mathematics is scored such that high scores correspond to negative attitudes towards Mathematics and vice versa.

We may rewrite $\tau_{XYc}^{\sim}(M_i)$ as

$$1 - E_I[var_I[h_c(X_{ij},Y_{ij})|\theta_i]|M_i]/var_I[h_c(X_{ij},Y_{ij})]$$

It follows that, given B, a sufficient condition for the meff of $T_{XYc}^{\sim}$ to have the simpler form of (5.92) is that the expectation of $var_I[h_c(X_{ij},Y_{ij})|\theta_i]$ does not depend on M_i. A diagnostic check of this condition is obtained by plotting $\hat{\sigma}_{c\,i}^2$ against M_i, where $\hat{\sigma}_{ci}^2$ is defined as in (5.49). Plots of $\hat{\sigma}_{ci}^2 = \hat{V}((X_{ij}-\mu_X)(Y_{ij}-\mu_Y)|\theta_i)$ against M_i are given in Figures 5.36-5.38 for the three pairs of variables from the National Survey of Attainment data. In these Figures there is little visual evidence of the regression functions depending on M_i. There may be a slight increase in the regression function in Figure 5.36 and a slight variation in Figure 5.37 but it certainly does not appear that $E(\hat{\sigma}_{ci}^2|M_i)$ increases montonically to an asymptote, as we would expect if the intra-cluster correlation were a decreasing function of M_i.

Figure 5.33- National Survey

X = Attitude to Mathematics Y = Mathematics Test

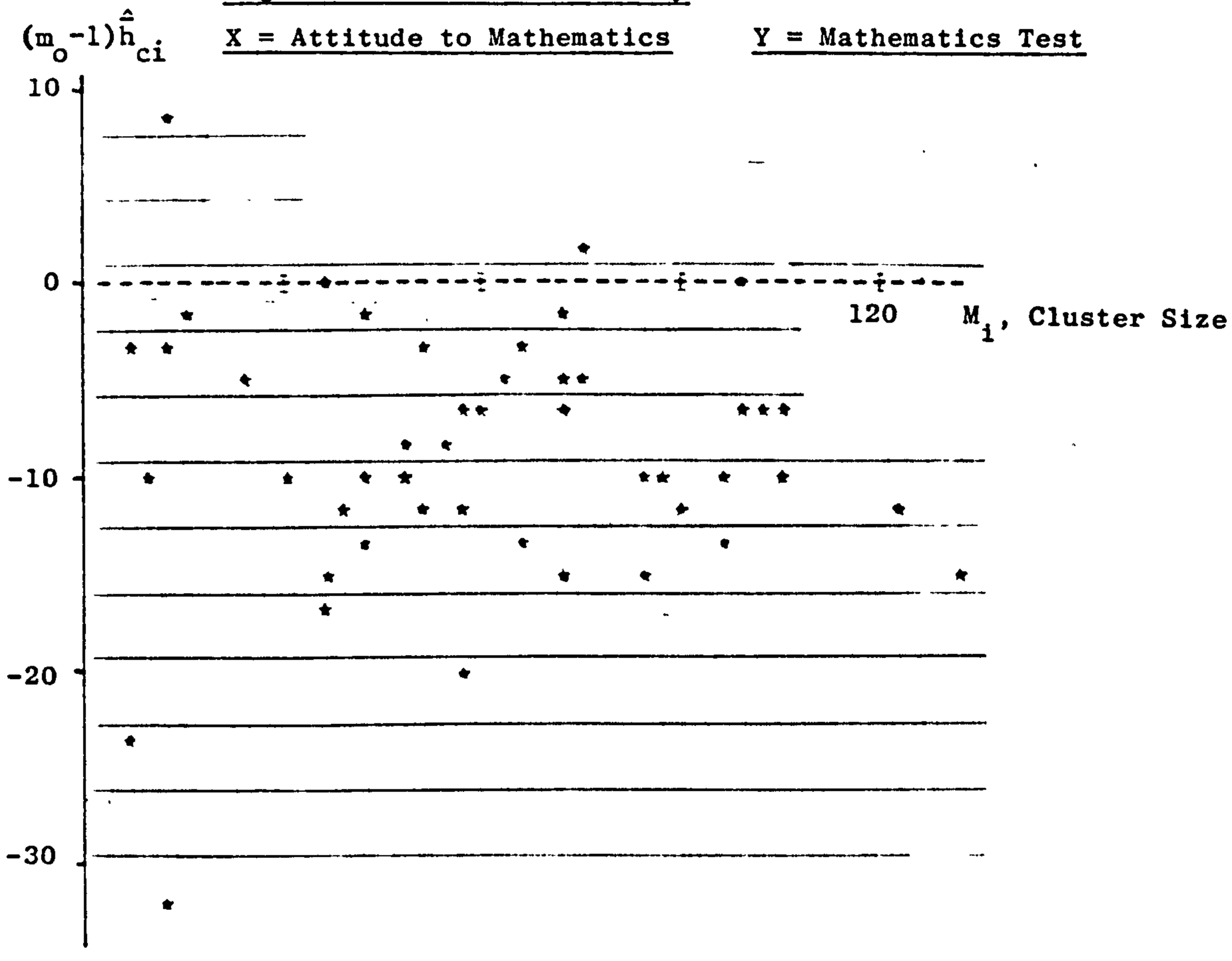


Figure 5.34- National Survey X = Attitude to Mathematics

Y = Welsh Reading Test

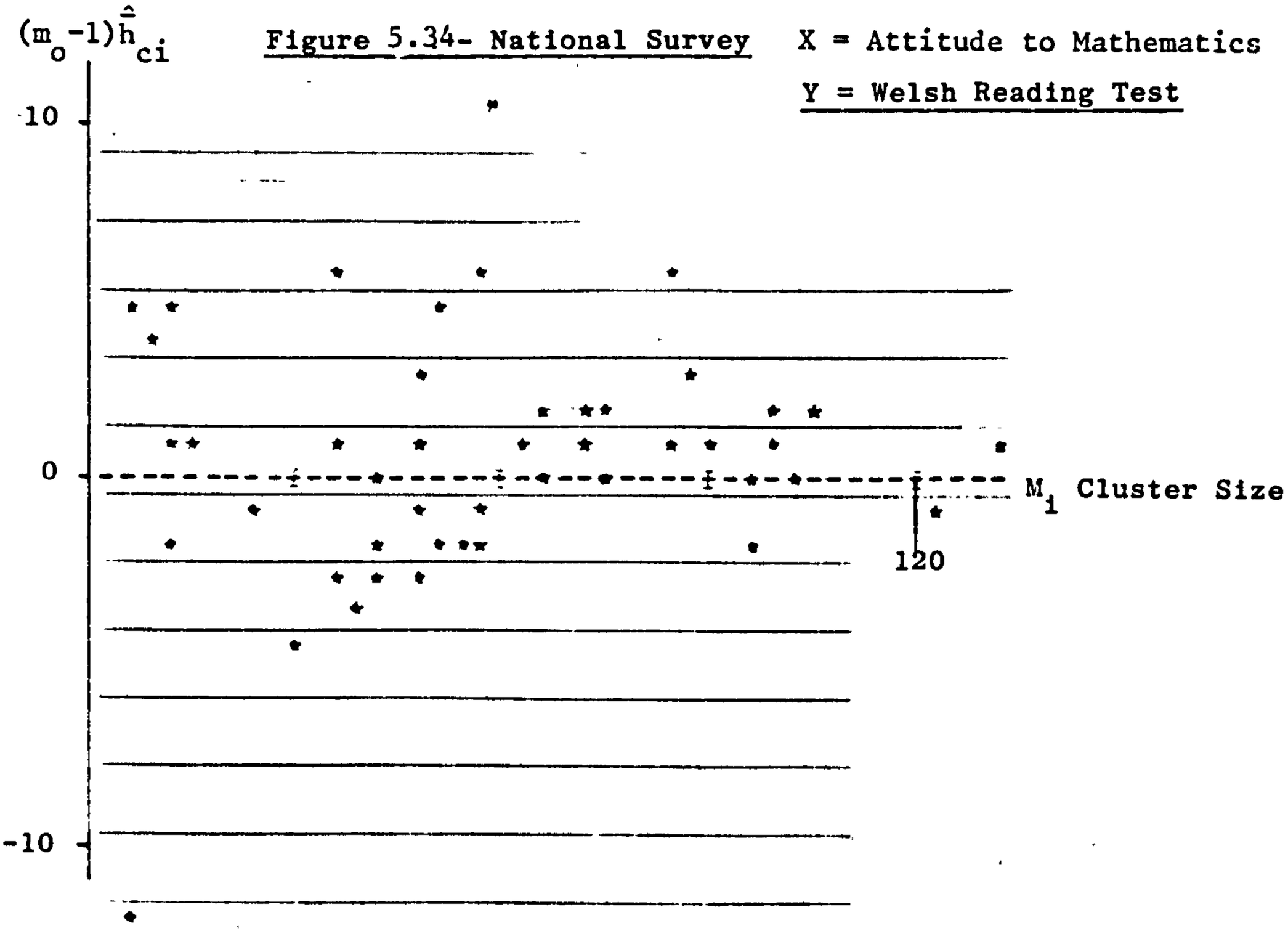


Figure 5.35- National Survey

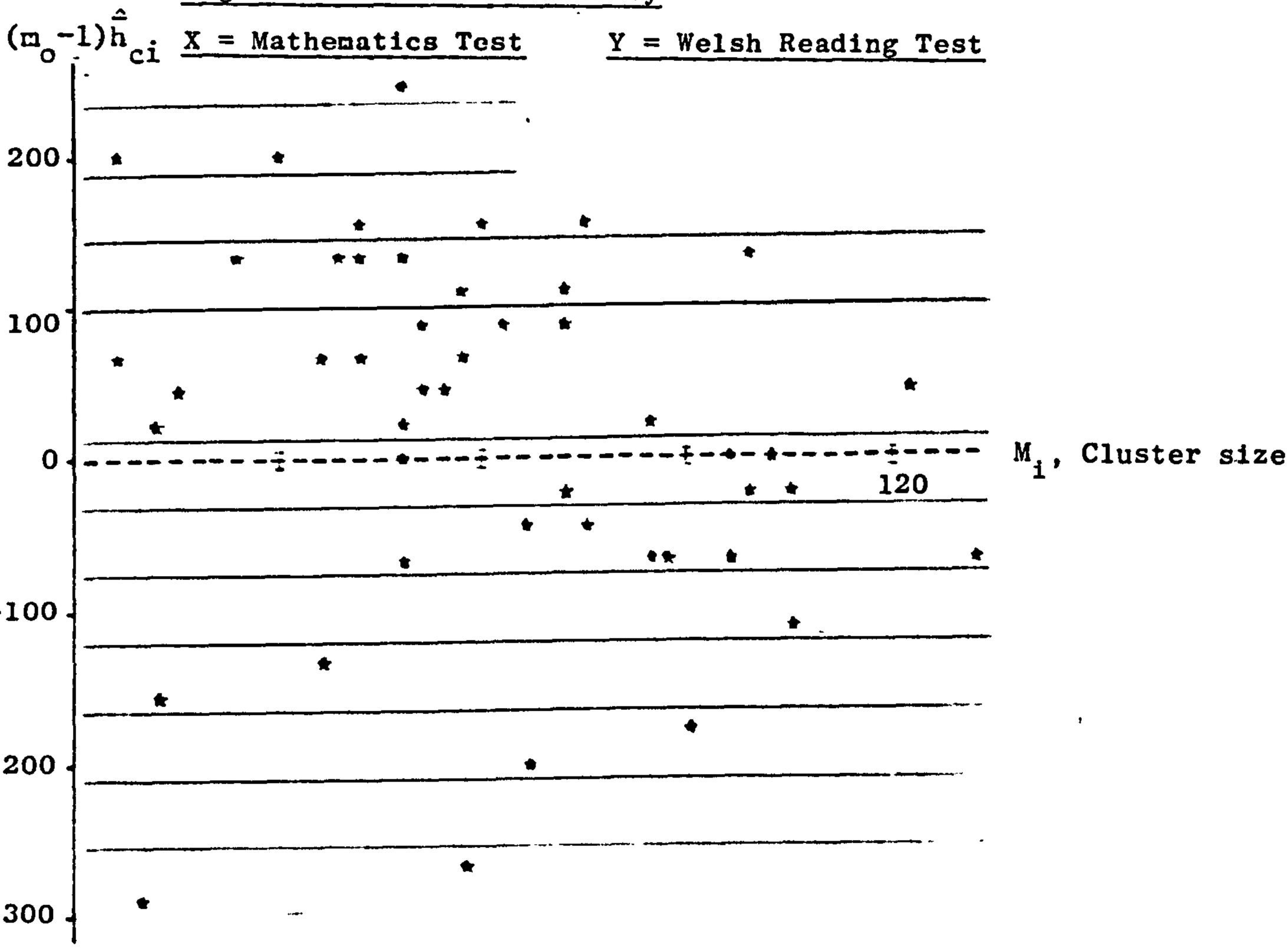


Figure 5.36 - National Survey

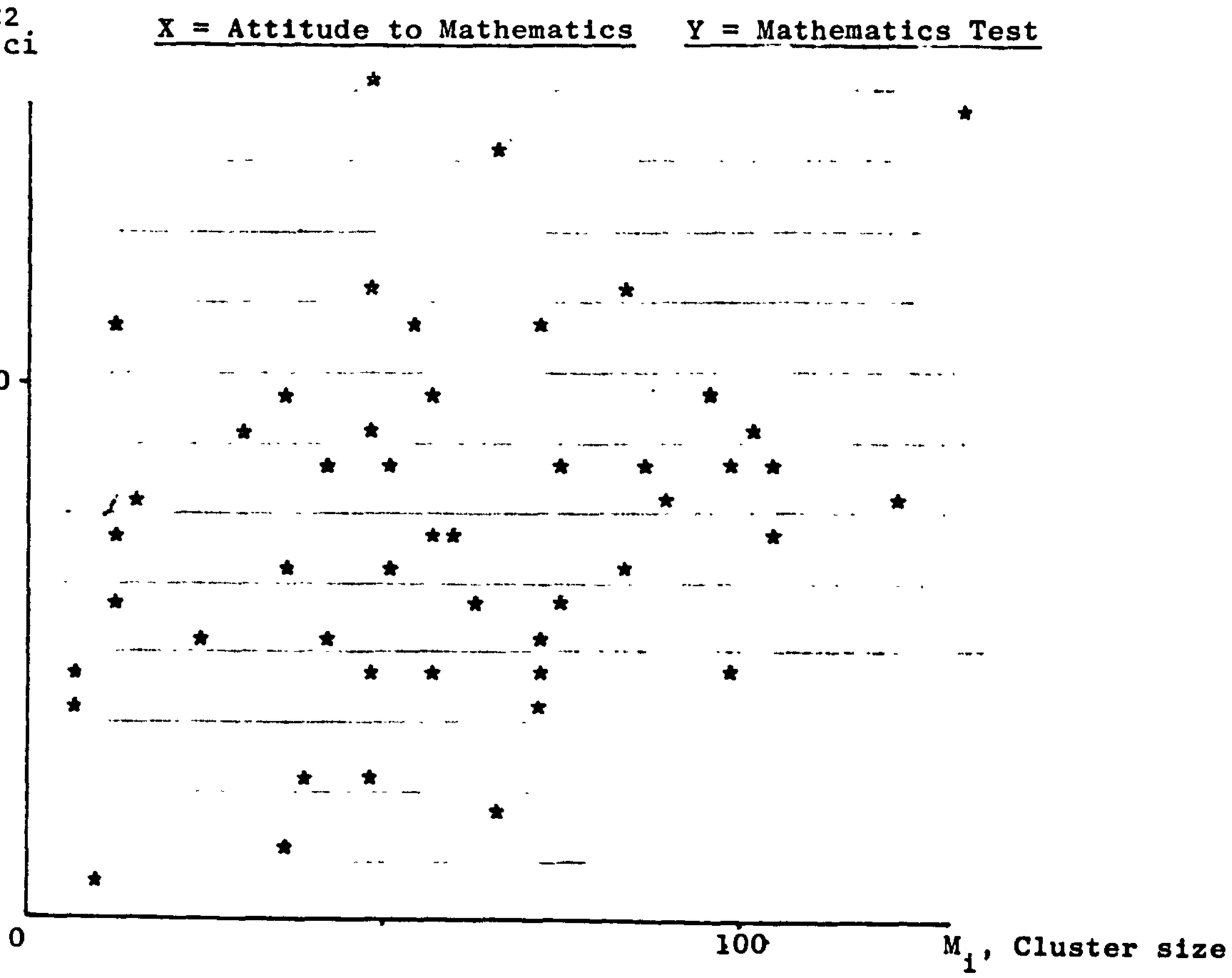


Figure 5.37- National Survey

X = Attitude to Mathematics Y = Welsh Reading Test

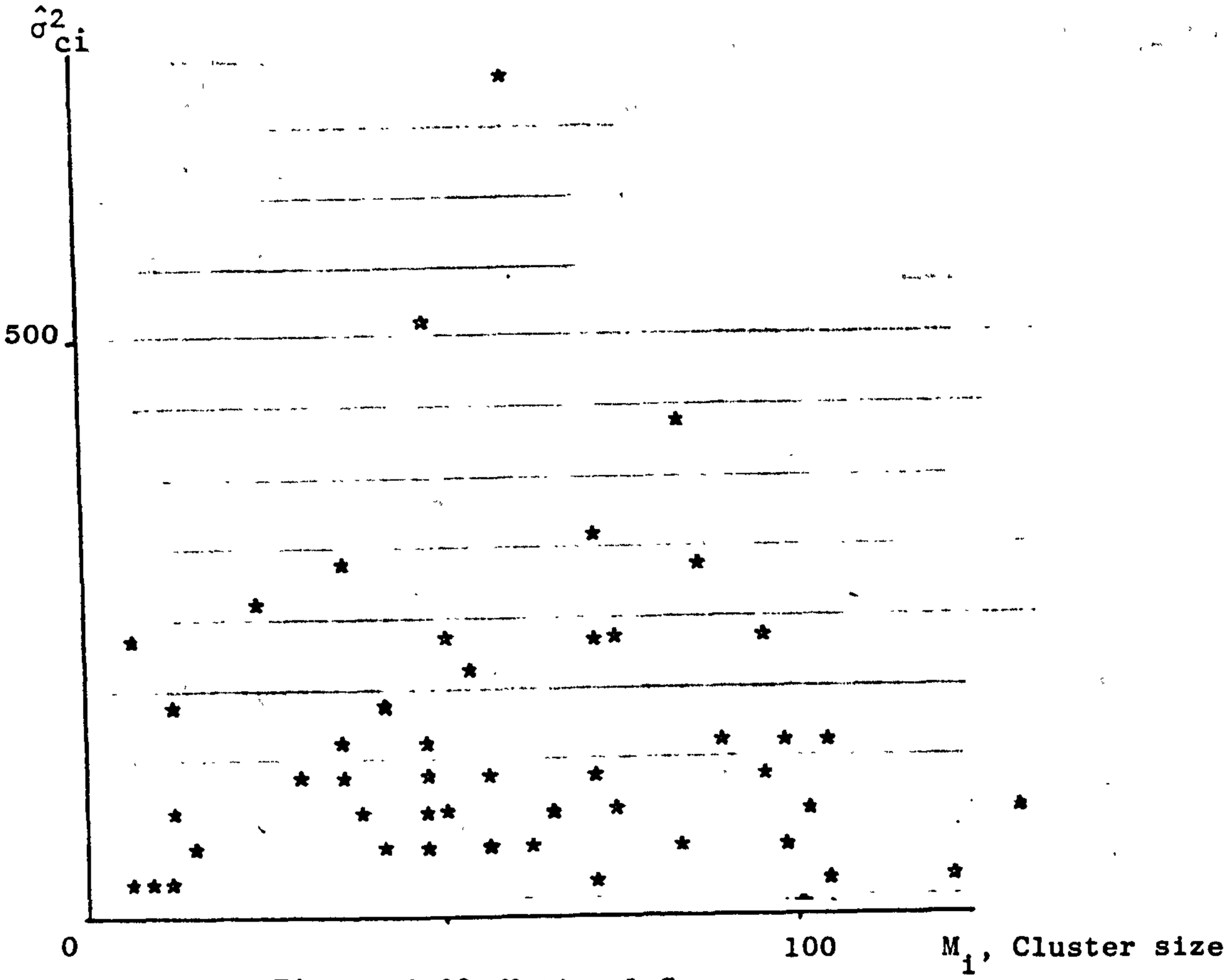
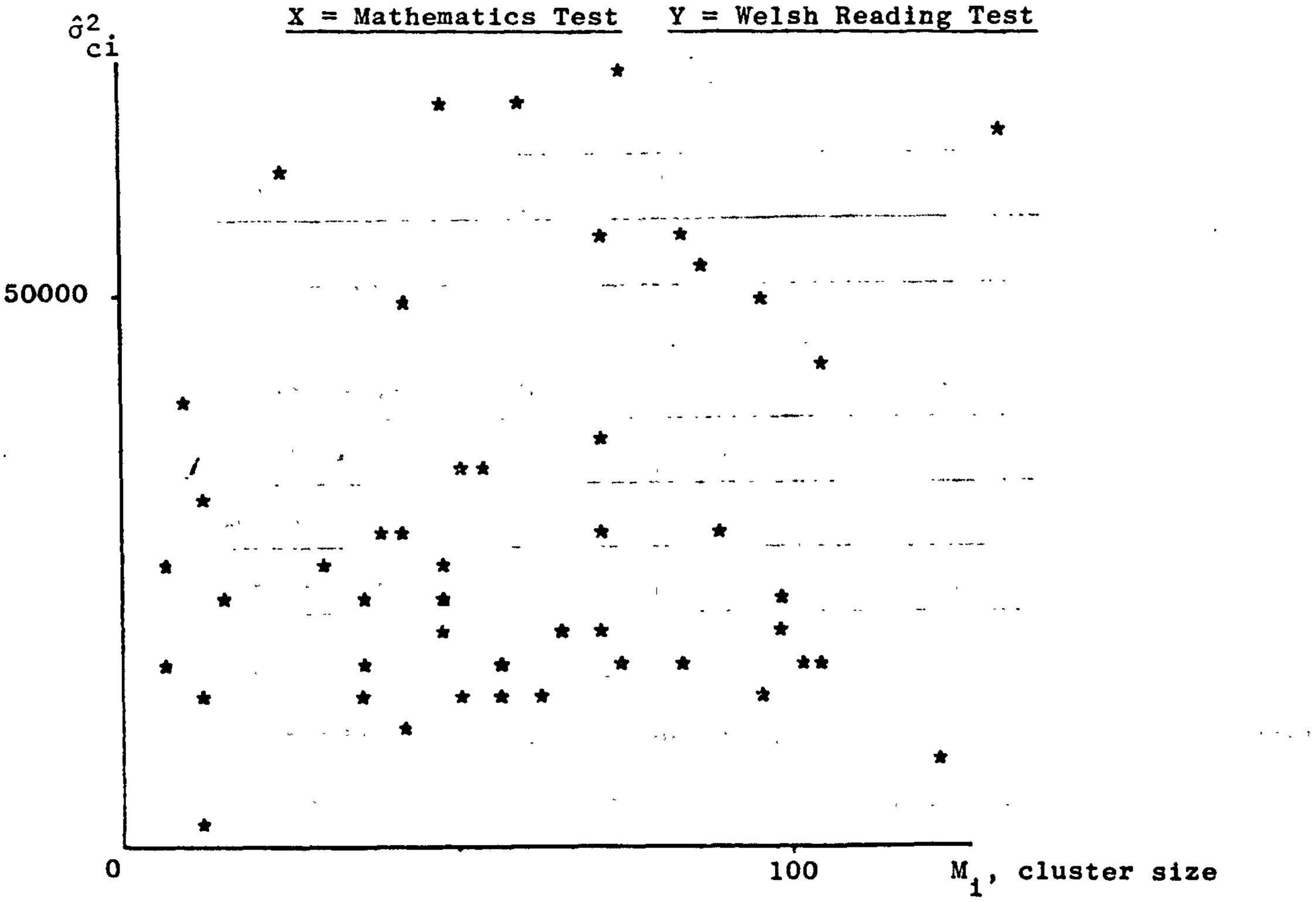


Figure 5.38- National Survey

X = Mathematics Test Y = Welsh Reading Test



We now compare the meff of $T_{XYc}^{\sim}$ with the meffs of T_{Xm} and T_{Ym}. We consider four special cases.

Case 1: A holds, $\sigma_{XYi} = \sigma_{XYW}$

If A holds

$$\tau_{XYc}^{\sim}(M_i) = \tau_{XYc}^{\sim} \qquad \text{(from corollary 5.26)}$$

If $\sigma_{XYi} = \sigma_{XYW}$ then

$$\gamma_{XY} = 0$$

$$c_{1XY} = 0$$

Hence $\tau_{XYc}^{\sim} = (\sigma_{XB}^2 \sigma_{YB}^2 + \sigma_{XYB}^2 + k_{22B})/(\sigma_X^2 \sigma_Y^2 + \sigma_{XY}^2 + k_{22})$

Let the overall correlation between X and Y, be $\rho = \sigma_{XY}/\sigma_X\sigma_Y$ and let the correlation between μ_{Xi} and μ_{Yi} be

$$\rho_B = \sigma_{XYB}/\sigma_{XB}\,\sigma_{YB} \qquad (5.95)$$

Then

$$\tau_{XYc}^{\sim} = \tau_{XM}\,\tau_{YM}\left(\frac{1+\rho_B^2 + K_{22B}}{1+\rho^2 + K_{22}}\right) \qquad (5.96)$$

where $K_{22B} = k_{22B}/\sigma_{XB}^2\,\sigma_{YB}^2$

$$K_{22} = k_{22}/\sigma_X^2\,\sigma_Y^2$$

Hence, assuming that the kurtoses K_{22B} and K_{22} do not differ greatly, $\tau_{XYc}^{\sim}$ will be small if either τ_{XM} or τ_{YM} is small and it will typically be smaller than both. This result is in the spirit of the conjectures of Kish and Frankel (1974). We illustrate the result by an example.

Example 5.10

Consider a population consisting of a mixture according to equal proportions, of five types of clusters within which X_{ij} is marginally

uniformly distributed on (0.9, 1.1), (1.9,2.1), (2.9,3.1), (3.9,4.1) and (4.9,5.1) respectively and Y_{ij} is marginally uniformly distributed on (0,2), (1,3), (2,4), (3,5) and (4,6) respectively (as in Example 5.6). Hence X is a highly clustered variable whereas the clustering on Y is somewhat less. We make no assumption about the joint distributions except that all within-cluster combinations of X_{ij} and Y_{ij} are possible. 100% probability regions for X_{ij} and X_{ij}' ($j \neq j'$) are plotted in Figure 5.39 and for Y_{ij} and Y_{ij}' in Figure 5.40 Corresponding regions are plotted in Figure 5.41 for $(X_{ij}-\mu_X)(Y_{ij}-\mu_Y)$ and $(X_{ij}'-\mu_X)(Y_{ij}'-\mu_Y)$ ($\mu_X=\mu_Y=3$). The Figures appear to confirm what we might expect from (5.96), that $\tau_{XY\tilde{c}}$ is less than both τ_{XM} and τ_{YM}.

Case 2: No clustering on one or both variables

If $\mu_{Xi} = \mu_X$ or $\mu_{Yi} = \mu_Y$ (alternatively if $\tau_{Xm} = 0$ or $\tau_{Ym} = 0$) then from (5.93)

$$\tau_{XY\tilde{c}}(M_i) = \gamma_{XY}(M_i)/(\sigma_X^2\sigma_Y^2 + \sigma_{XY}^2 + k_{22}) \qquad (5.97)$$

Hence the misspecification effect depends fundamentally on the variation in covariances between clusters. In particular, even if both X and Y exhibit no marginal clustering it is not necessary that $\tau_{XY\tilde{c}}(M_i) = 0$ (as in Case 1). To illustrate this point we give an example.

Example 5.11

Consider a population consisting of a mixture, in equal proportions of two types of cluster. In both types of cluster (X_{ij}, Y_{ij}) has a bivariate normal distribution with zero means and unit variances. In the first type of cluster $\sigma_{XYi} = \rho$ and in the second type $\sigma_{XYi} = -\rho$. Plots of within-cluster concentration eclipses are given in Figure 5.42

Figure 5.39- Example 5.10
100% probability
regions for X and X'

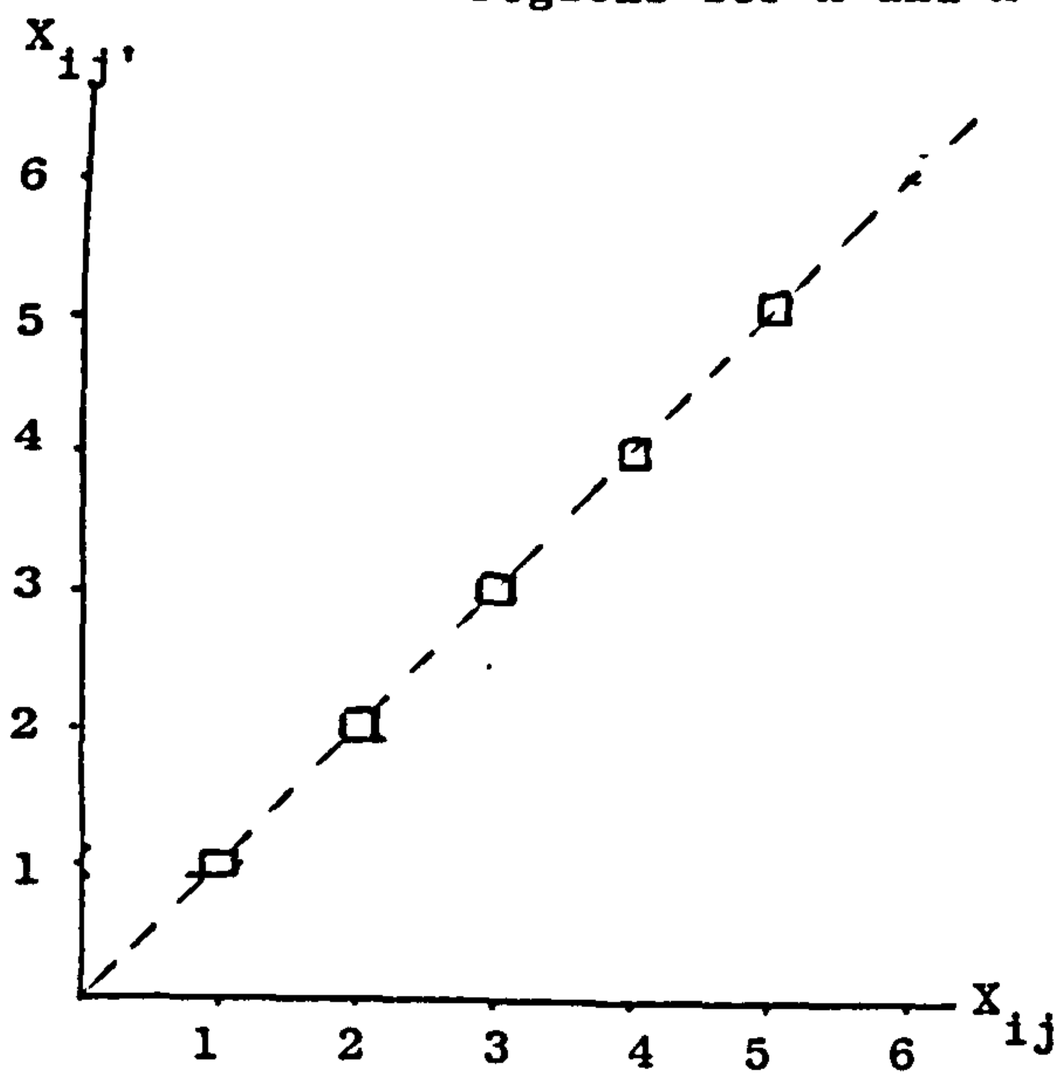


Figure 5.40- Example 5.10
100% probability region
for Y and Y'

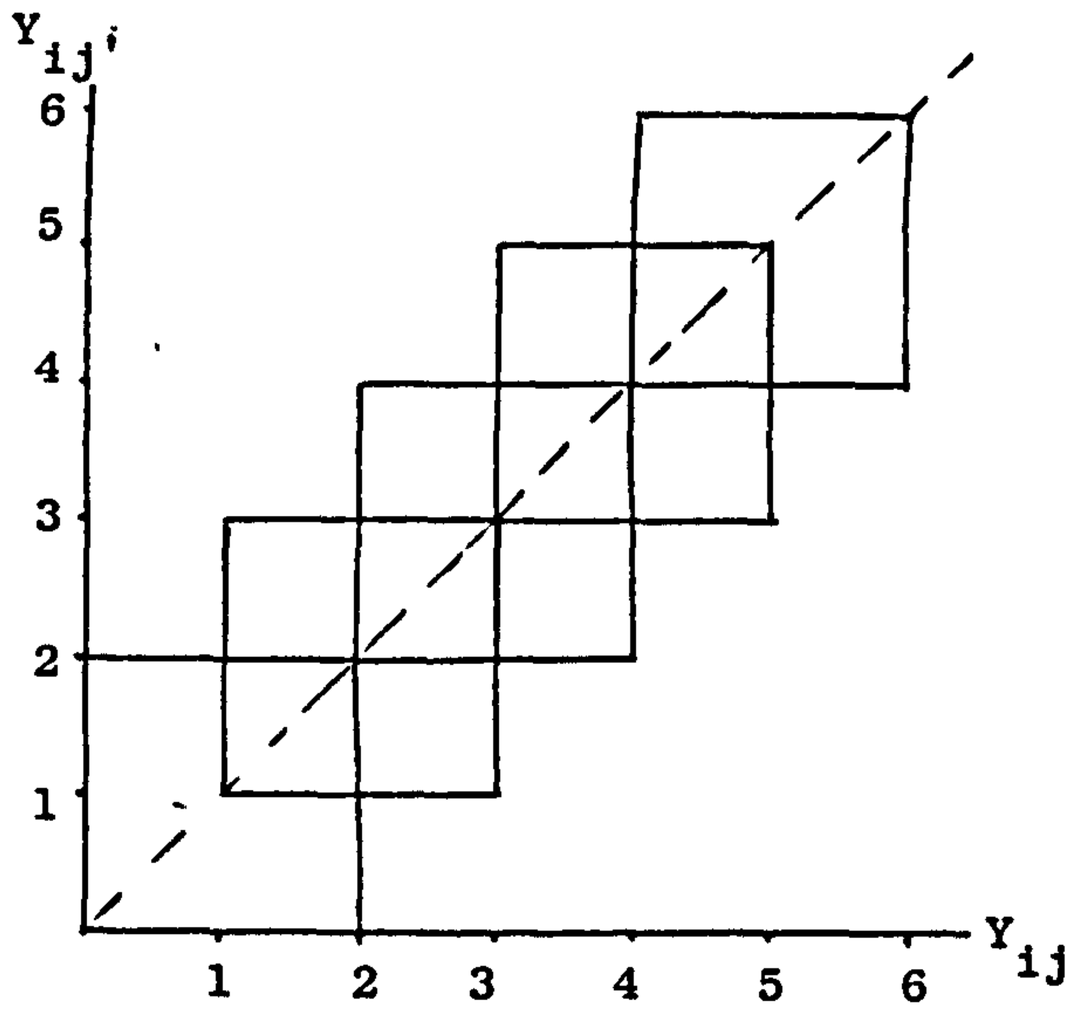


Figure 5.41- Example 5.10
100% probability region for
$(X-\mu_x)(Y-\mu_y)$ and $(X'-\mu_x)(Y'-\mu_y)$

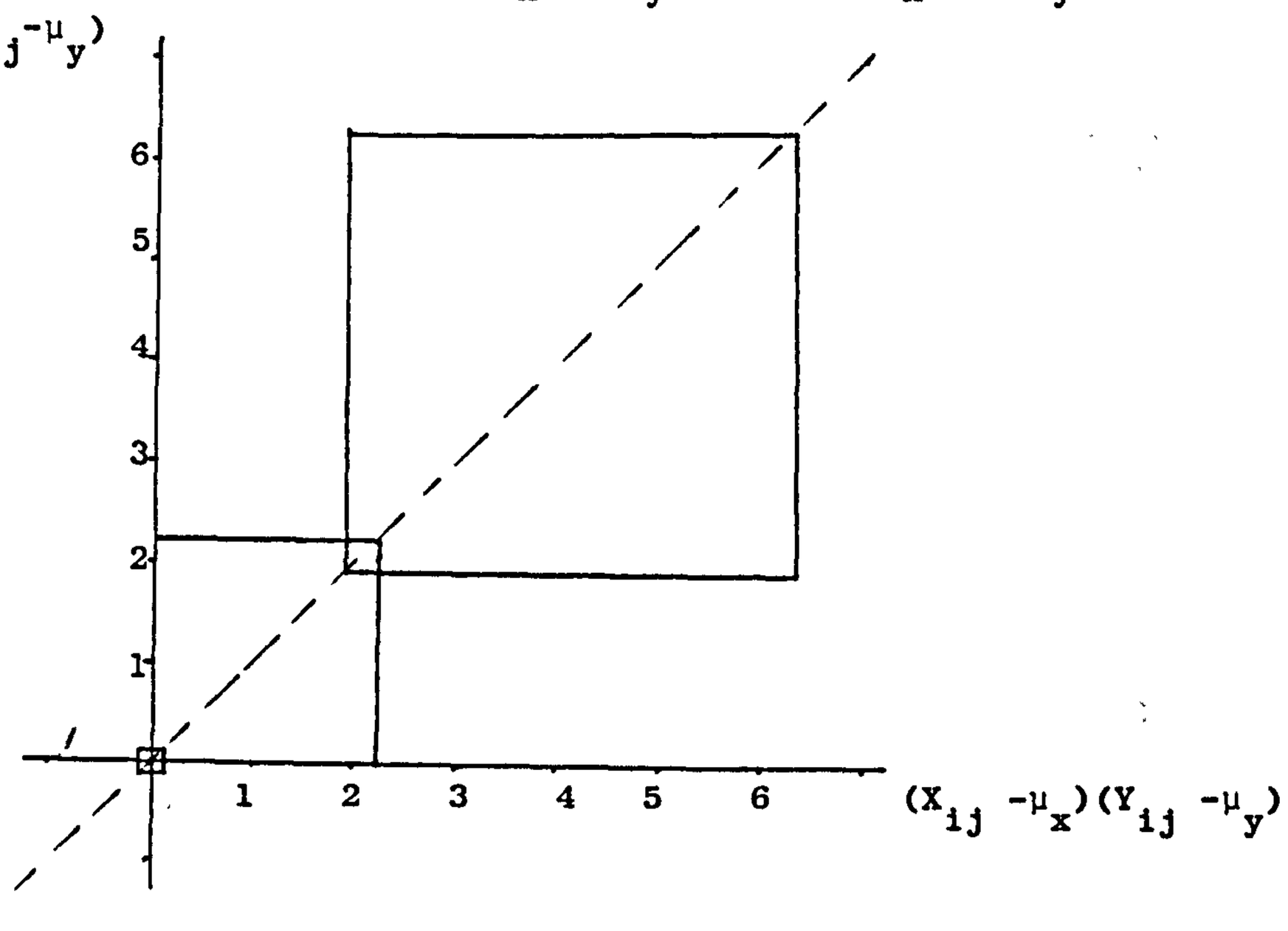


Figure 5.42 - Example 5.11
Within-cluster concentration eclipses

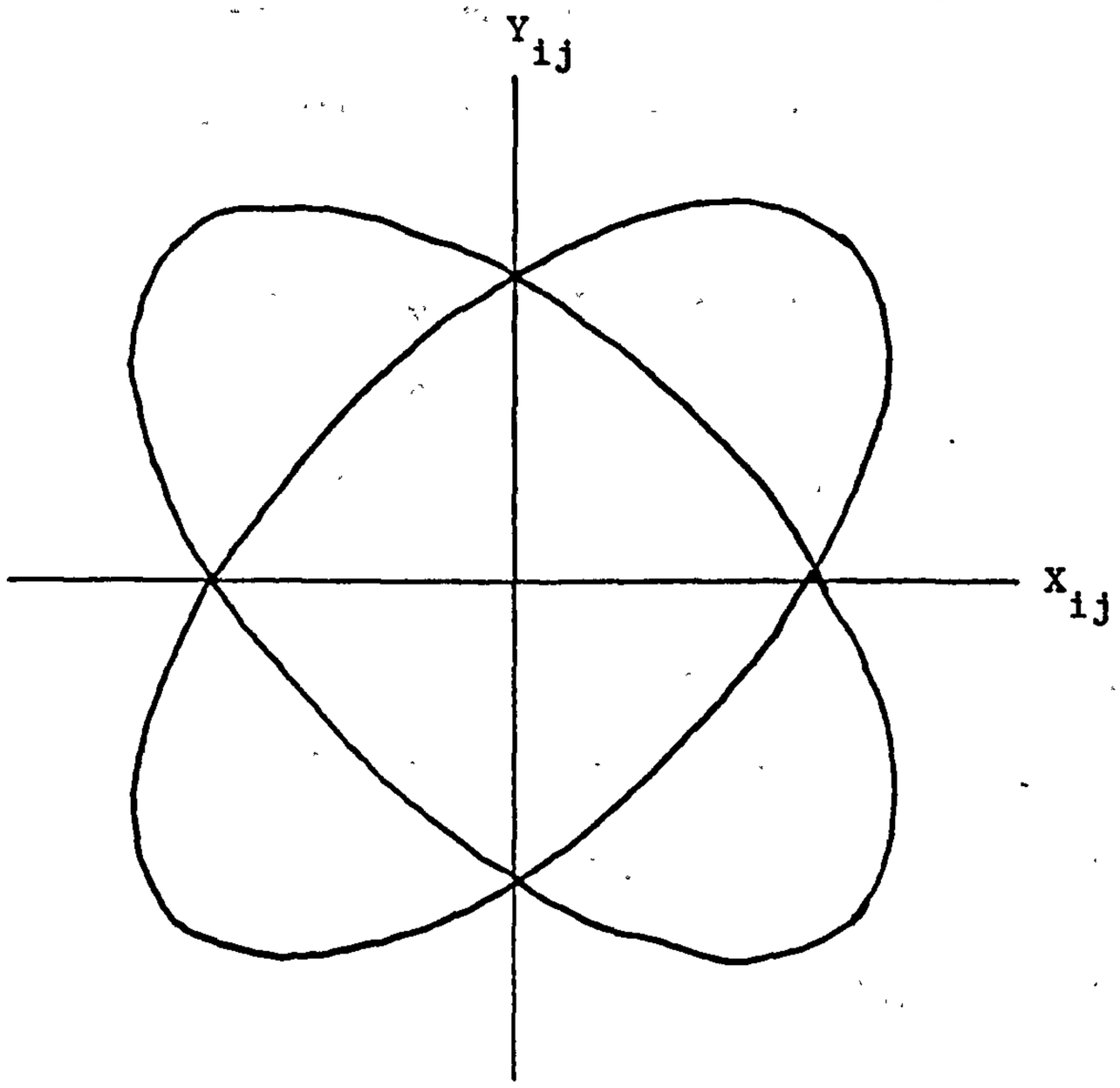
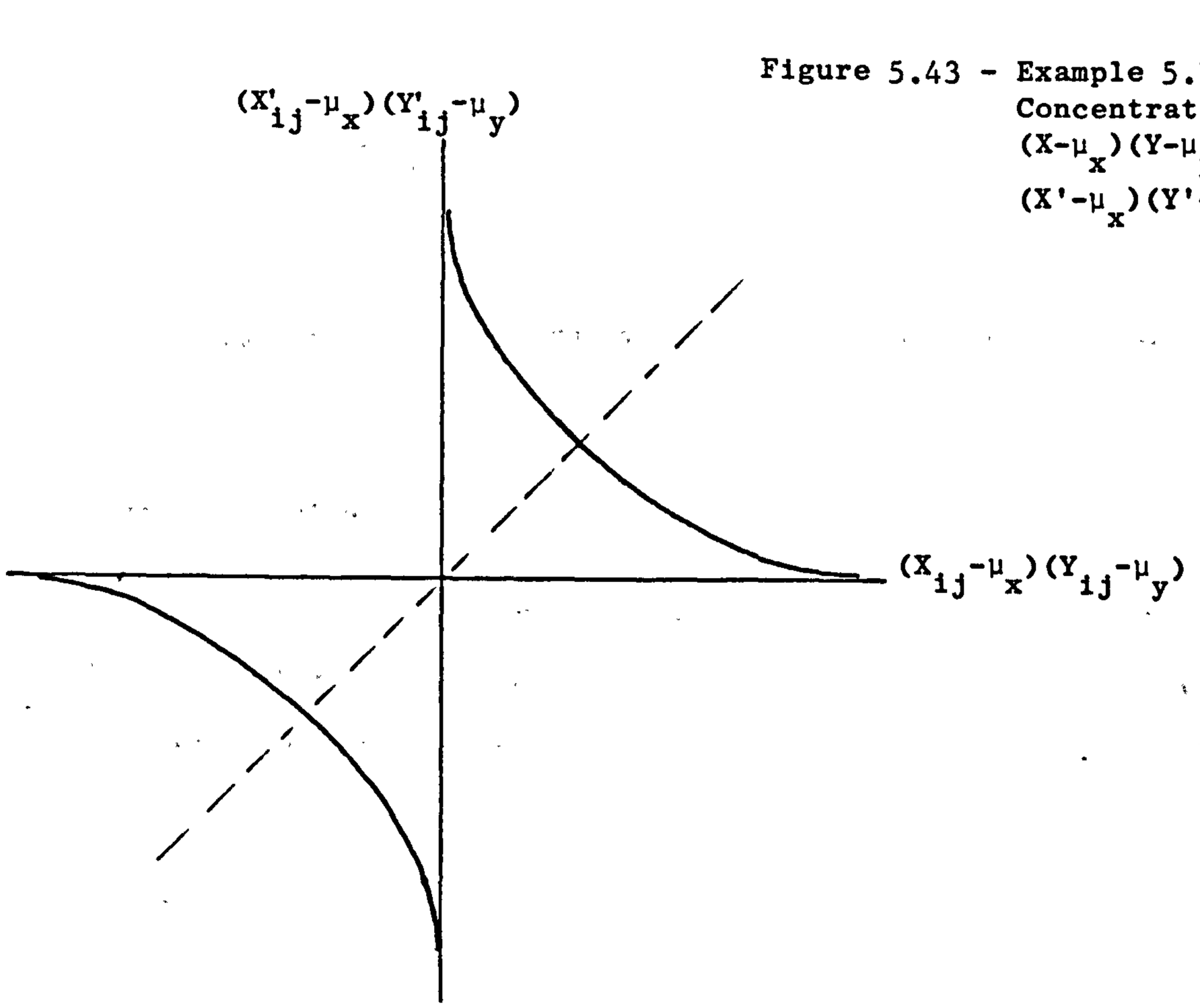


Figure 5.43 - Example 5.11
Concentration regions for $(X-\mu_x)(Y-\mu_y)$ and $(X'-\mu_x)(Y'-\mu_y)$ when $\rho=1$



Within-cluster probability regions for X_{ij} and X_{ij}' ($j \neq j'$) or Y_{ij} and Y_{ij}' are just circles of the same diameter, centre (0,0), since both X and Y exhibit no marginal clustering. However, if we plot $(X_{ij}-\mu_X)(Y_{ij}-\mu_Y)$ against $(X_{ij}'-\mu_X)(Y_{ij}'-\mu_Y)$ ($j \neq j'$, $\mu_X = \mu_Y = 0$) we find that, assuming $\rho > 0$, the first type of cluster tends to be concentrated in the positive quadrant while the second type of cluster is concentrated in the negative quadrant, making $\tau_{XYc}\tilde{}(M_i)$ non-zero. The extreme case where $\rho = 1$ is plotted in Figure 5.43. In this example (5.97) simplifies to

$$\tau_{XYc}\tilde{} = \gamma_{XY}/(1+2\gamma_{XY}) = \rho^2/(1+2\rho^2)$$

since $k_{22} = E_I[E_I[(X_{ij}-\mu_{Xi})^2(Y_{ij}-\mu_{Yi}^2)|\theta_i]|M_i] - \sigma_X^2\sigma_Y^2 - \sigma_{XY}^2$

$$= E_I[\sigma_{Xi}^2\sigma_{Yi}^2 + 2\sigma_{XYi}^2|M_i] - \sigma_X^2\sigma_Y^2 - \sigma_{XY}^2$$

$$= 2\gamma_{XY}$$

and

$$\gamma_{XY} = \text{var}(\sigma_{XYi})$$

$$= \rho^2/2 + \rho^2/2 = \rho^2$$

Hence $\tau_{XYc}\tilde{}$ is bounded above by 1/3, and is equal to this value when $\rho = \pm 1$.

Case 3: A holds

In general, if A holds, $\tau_{XYc}\tilde{}(M_i) = \tau_{XYc}\tilde{}$ will be a combination of the expressions (5.96) and (5.97). If (μ_{Xi}, μ_{Yi}) are normally distributed between clusters we may write from (5.94)

$$\tau_{XYc}\tilde{} = \tau_1 + \tau_2 + \tau_3$$

where

$$\tau_1 = \tau_{Xm}\tau_{Ym}(1+\rho_B^2)\ \sigma_X^2\sigma_Y^2/(\sigma_X^2\sigma_Y^2(1+\rho^2) + k_{22})$$

$$\tau_2 = 2c_{1XY}/(\sigma_X^2\ \sigma_Y^2(1+\rho^2) + k_{22})$$

$$\tau_3 = \gamma_{XY}/(\sigma_X^2\sigma_Y^2(1+\rho^2) + k_{22})$$

Some estimates of these quantities for the three Family Expenditure Survey variables are given in Table 5.4. The method of estimation is described in the Appendix.

Table 5.4 Estimates for FES Data

Variables X,Y	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_{XY\tilde{c}}$
log(V1),log(V2)	.0006	-.0012	.0154	.0148
log(V1),log(V3)	.0012	-.0014	.0075	.0073
log(V2),log(V3)	.0005	-.0010	.0164	.0159

It is clear from Table 5.4 that for each pair of variables $\hat{\tau}_{XY\tilde{c}}$ is dominated by $\hat{\tau}_3$. This is simply explained. τ_1 is very small in each case because the τ_{Xm}'s are small (see Table 5.2) and the overall kurtoses k_{22} are not far from zero. Furthermore, a simple consequence of the definition of c_{1XY} is that

$$|\tau_2| \leq 2(\tau_1\tau_3)^{\frac{1}{2}}$$

and hence if τ_1 is very small then so is τ_2. Hence $\tau_{XY\tilde{c}}$ is mainly determined by τ_3. τ_3 might be taken as a standardised measure of the variation in cluster covariances, σ_{XYi}.

Case 4: B holds but A does not hold

As in Section 5.4, we just consider the spatial process approach. We now assume that (X,Y) is a stationary isotropic bivariate spatial process and from (5.56)

$$\tau_{XY\tilde{c}}(M_i) = \int_o^1 \rho_{hc}(\alpha_i t)K(t)dt$$

where $\rho_{hc}(s) = \text{corr}\left[h_c(X(\underline{x}), Y(\underline{x})), h_c(X(\underline{x}'), Y(\underline{x}')\right]$

and s is the distance between $\underline{x}$ and $\underline{x}'$.

Let $\gamma_X(s) = \text{cov}\left[X(\underline{x}), X(\underline{x}')\right]$

$\gamma_Y(s) = \text{cov}\left[Y(\underline{x}), Y(\underline{x}')\right]$

$\gamma_{XY}(s) = \text{cov}\left[X(\underline{x}), Y(\underline{x}')\right]$

$\gamma_{YX}(s) = \text{cov}\left[Y(\underline{x}), X(\underline{x}')\right]$

We assume $\gamma_{XY}(s) = \gamma_{YX}(s)$

Let $r_X(s) = \gamma_X(s)/\sigma_X^2$, $r_Y(s) = \gamma_Y(s)/\sigma_Y^2$, $r_{XY}(s) = \gamma_{XY}(S)/\sigma_X\sigma_Y$

Then $r_X(s)$ and $r_Y(s)$ correspond to $\rho(s)$ in (5.57). We shall assume that the process in Gaussian, in which case

$$\rho_{hc}(s) = \text{corr}\left[(X(\underline{x})-\mu_X)(Y(\underline{x})-\mu_Y) , (X(\underline{x}')-\mu_X)(Y(\underline{x}')-\mu_Y)\right]$$

$$= \frac{\gamma_X(s)\gamma_Y(s) + \gamma_{XY}^2(s)}{\sigma_X^2\,\sigma_Y^2 + \sigma_{XY}^2}$$

$$= \frac{r_X(s)\; r_Y(s) + r_{XY}\;^2(s)}{1 + \rho^2}$$

where $\rho = \sigma_{XY}/\sigma_X\sigma_Y$ as before.

Hence

$$\tau_{XY\tilde{c}}(M_i) = \int_o^1 \left[r_X(\alpha_i t)r_Y(\alpha_i t) + r_{XY}^2(\alpha_i t)\right]K(t)dt/(1+\rho^2) \qquad (5.98)$$

Now from Schwarz's inequality

$$\int_o^1 r_X(\alpha_i t)r_Y(\alpha_i t)K(t)dt \leq \left[\int_o^1 r_X 2(a_i t)K(t)dt \int_o^1 r_Y{}^2(\alpha_i t)K(t)dt\right]^{\frac{1}{2}}$$

$$= (\tau_{X\tilde{v}}(M_i)\tau_{Y\tilde{v}}(M_i))^{\frac{1}{2}}$$

$$\leq (\tau_{Xm}(M_i)\tau_{Ym}(M_i))^{\frac{1}{2}} \qquad \text{from (5.58)}$$

If we also make the assumption that

$$r_{XY}^2(s) \leq \rho^2 r_X(s)r_Y(s) ,$$

which would seem reasonable in practice, then it follows that

$$\tau_{XY\tilde{c}}(M_i) \leq \left(\tau_{Xm}(M_i)\tau_{Ym}(M_i)\right)^{\frac{1}{2}}$$

If there is little spatial correlation on one variable, say Y , so that $\tau_{Ym}(M_i)$ decays very quickly with respect of M_i then $\tau_{XY\tilde{c}}(M_i)$ will also decay very quickly with respect to M_i.

5.6 Conclusion

In this chapter we have considered the properties of estimators based on the assumption that observations are IID. Under a general clustered population model it was shown that misspecification of the model as IID does not introduce model-bias provided Assumption B holds, i.e. provided the marginal distributions of the observations within clusters do not depend on the cluster sizes. The true model-variance of the estimate is, however, greater than the IID-variance by a factor referred to as the misspecification effect (Definition 5.1). Similar results hold for the design-model-bias and variance even if B does not hold providing the design is self-weighting.

The form of this misspecification effect was investigated for means, variances and covaraiances. The misspecification effect depends on a statistic T only via a generalised intra-cluster correlation $\tau(T)$ (which may also depend on the cluster size). Conjectures given by Kish and Frankel (1974) that design effects for complex statistics are

less than those of means and formulae such as that in Bebbington and Smith (1977, p.185) relating design effects for complex statistics to intra-cluster correlations for means suggest that we might be able to relate τ (variance) and τ (covariance) to τ (mean). It turns out that no simple relations necessarily hold between these quantities. For example, the usual intracluster correlations for the means of X and Y might both be zero whereas the misspecification effects for the variances of X and Y or for the covariance between X and Y may all be greater than one. On the basis of algebraic arguments and empirical work on Family Expenditure Survey data we conjecture that τ (variance) will usually be largely determined by the variance between cluster variances and τ (covariance) by the variation between cluster covariances provided the overall intra-cluster correlations (on the means) are not excessive (say > 0.2). Such results correspond to the interpretation of τ (mean) as a measure of variation between cluster means.

We also argue that if the clusters may be viewed as regions of constant shape in an isotropic stationary spatial process of constant population density then τ (variance) and τ (covariance) will generally be less than the corresponding τ (mean)'s. Further empirical investigation such as that of Proctor (1980) is necessary in order to formulate functional relationships between the τ's and the cluster size. Such an approach might be useful for the design of analytical surveys. For example, for a given cost function (e.g. Brewer et. al,, 1977) it may be that when estimating covariances fewer clusters need be sampled than when estimating means to attain a given precision.

CHAPTER SIX - ALTERNATIVE ESTIMATORS UNDER TWO-STAGE SAMPLING

6.1 Introduction

In this chapter we assume that Model I of Section 5.1 is true and moreover that Assumption A (of the same section) is also true. Hence, in particular,

$$E_I(Y_{ij}|M_i) = \mu \tag{6.1}$$

$$V_I(Y_{ij}|M_i) = \Sigma \tag{6.2}$$

where μ and Σ are the mean vector and covariance matrix respectively of f_o defined in (5.3) and where we now take Y_{ij} to be a general pxl vector.

In Section 6.2 we shall consider the model-based estimation of μ and Σ. In Section 6.3 we consider model-based predictors of

$$\bar{y} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}/M_o \tag{6.3}$$

and

$$S = \sum_{i=1}^{N} \sum_{j=1}^{M_i} (y_{ij}-\bar{y})(y_{ij}-\bar{y})'/(M_o-1). \tag{6.4}$$

In Section 6.4 we consider design-based predictors of $\bar{y}$ and S.

As noted after Lemma 5.1, $(s,\underline{M})$ will be ancillary for μ and Σ if Assumption A holds. Hence our inference procedures are based on the (Model I) sampling distribution of the y_{ij} given s and $\underline{M}$.

It is somewhat unfortunate that our discussion is restricted to the case when Assumption A holds, since as noted in Chapter 5 the standard estimators will be least satisfactory when Assumption B (and hence A) does not hold. However, as will be seen, the 'optimal' estimation of μ and Σ even under Assumption A is not that easy and it seems necessary to deal with the simplest case first.

6.2 Model-Based Estimation

In Section 3.2 we only used one estimation method - maximum likelihood. In the variance components literature a number of other methods have also been used for various reasons (Harville, 1977) and we therefore consider in addition some of these methods. Our model is essentially a generalised multivariate random effects model where the random effects are

$$\mu_i = E_I(Y_{ij}|\theta_i) \tag{6.5}$$

and

$$\Sigma_i = V_I(Y_{ij}|\theta_i) \tag{6.6}$$

This seems to us a more natural model than models which take μ_i as random and Σ_i as fixed (e.g. Rao et al, 1981). The estimation problem for our model is not, however, particularly easy and so we consider progressively more general cases, viz. Case 1 : $m_i=m$, $\Sigma_i=\Sigma_W$, Case 2 : m_i unequal and $\Sigma_i=\Sigma_W$ and Case 3 : m_i unequal, Σ_i unequal.

Case 1 : $m_i=m$, $\Sigma_i=\Sigma_W$

This is the conventional balanced one-way multivariate random effects model (e.g. Searle, 1956). The basic parameters are μ $\Sigma_W = V_I(Y_{ij}|\theta_i)$ and $\Sigma_B = V_I(\mu_i|M_i)$.

ANOVA Estimation

Lemma 6.1

The ANOVA estimators are

$$\tilde{\mu}_{ANOVA} = \bar{y}_s = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}/mn \tag{6.7}$$

$$\hat{\Sigma}_{ANOVA} = S_s^B/M + (m-1)\, S_s^W/m \tag{6.8}$$

where

$$S_s^B = m \sum_{i=1}^{n} (\bar{y}_i-\bar{y}_s)(\bar{y}_i-\bar{y}_s)'/(n-1) \tag{6.9}$$

$$S_s^W = \sum_{i=1}^{n}\sum_{j=1}^{m} (y_{ij}-\bar{y}_i)(y_{ij}-\bar{y}_i)'/n(m-1) \tag{6.10}$$

$$\bar{y}_i = \sum_{j=1}^{m} y_{ij}/m \qquad (6.11)$$

$\hat{\mu}_{ANOVA}$ and $\hat{\Sigma}_{ANOVA}$ are unbiased for μ and Σ respectively given s and $\underline{M}$.

Proof : Let $\tilde{y} = (y_{11}...y_{1m}y_{21}...y_{nm})'$ be the nmxp data matrix. Let $L_m = 1_{nm}'/\sqrt{mn}$, where 1_{nm} is the mnx1 vector of ones. Let L_B be a (n-1) xnm matrix such that the rows of $\begin{pmatrix} L_m \\ L_B \end{pmatrix}$ form an orthonormal basis of the subspace of R^{mn} spanned by vectors e_i (i=1...n) which have ones in the $[i(m-1) + 1]^{th}$ to im^{th} positions and zeros elsewhere. Let L_W be the (nm-n) xnm matrix such that the rows of

$$L = \begin{pmatrix} L_m \\ L_B \\ L_W \end{pmatrix}$$

form an orthonormal basis of R^{mn}.

Hence $$LL' = L'L = I_{mn}$$

The ANOVA decomposition is then

$$\tilde{y}'\tilde{y} = \tilde{y}'L'L\tilde{y} = \tilde{y}'L_m'L_m\tilde{y} + \tilde{y}'L_B'L_B\tilde{y} + \tilde{y}'L_WL_W\tilde{y}$$

and $$\tilde{y}'L_m'L_m\tilde{y} = mn\ \bar{y}_s\bar{y}_s'$$

$$\tilde{y}'L_B'L_B\tilde{y} = (n-1)\ S_s^B$$

$$\tilde{y}'L_W'L_W\tilde{y} = n(m-1)S_s^W$$

The 'ANOVA' estimator of μ is $\bar{y}_s = L_m\tilde{y}/\sqrt{mn}$ and is unbiased for μ since

$$E_I(\tilde{y}|\underline{M},S) = 1_{mn}\mu'$$

where we ignore the distinction between y_{ij} and Y_{ij}. To obtain the ANOVA estimators of Σ_B and Σ_W write $\tilde{y}$ in the linear model form

$$\tilde{y} = 1_{mn}\ \mu' + diag_n(1_m)\ \tilde{a} + \tilde{e} \qquad (6.12)$$

where $\tilde{a}' = [(\mu_1-\mu)' \ ... \ (\mu_n-\mu)']'$

$\tilde{e}' = [(y_{11}-\mu_1)' \ ... \ (y_{1m}-\mu_1)' \ ... \ (y_{nm}-\mu_n)']'$

Note that $L_W\tilde{y} = L_W\tilde{e}$ and $E_I(\tilde{e}\,\tilde{e}'|\underline{M},s) = \Sigma_W$ so that

$$E_I(\tilde{y}'\,L_W'\,L_W\,\tilde{y}|s,\underline{M}) = E_I(\tilde{e}'L_W'L_W\tilde{e}|s,\underline{M})$$

$$= tr(L_W'L_W)E_I(\tilde{e}\tilde{e}'|s,\underline{M})$$

$$= tr(L_W'L_W)\;\Sigma_W$$

Also by definition $L_W'\,L_W = diag_n(P_{Bm})$ where $P_{Bm} = I_m - 1_m 1_m'/m$ so that

$$E_I(\tilde{y}'L_W'L_W\tilde{y}|s,\underline{M}) = ntr(P_{Bm})\Sigma_W$$

$$= n(m-1)\Sigma_W$$

Hence $E_I(S_s^W|s,\underline{M}) = \Sigma_W$ and S_s^W is the unbiased ANOVA estimate of Σ_W.

Similarly

$$E_I(\tilde{y}'L_BL_B\tilde{y}|s,\underline{M}) = E_I(\tilde{a}'mP_{Bn}\tilde{a} + \tilde{e}'L_B'L_B\tilde{e}|s,\underline{M})$$

$$= tr(mP_{Bn})\Sigma_B + tr(L_B'L_B)\Sigma_W$$

$$= m(n-1)\;\Sigma_B + (n-1)\Sigma_W$$

$$(L_B'L_B = diag(1_m 1_m'/m) - 1_{mn}1_{mn}'/mn)$$

Hence $(S_s^B - S_s^W)/m$ is the unbiased ANOVA estimate of Σ_B. Hence the unbiased ANOVA estimator of $\Sigma = \Sigma_B + \Sigma_W$ is

$$\hat{\Sigma}_{ANOVA} = (S_s^B - S_s^W)/m + S_s^W$$

$$= S_s^B/m + (m-1)\;S_s^W/m$$

Lemma 6.2

$$\text{If } Y_{ij}|\theta_i \sim N_p(\mu_i,\ \Sigma_W)$$

$$\text{and } \mu_i|M_i \sim N_p(\mu,\ \Sigma_B) \qquad (6.13)$$

then $$V_I(\hat{\mu}_{ANOVA}|s,\underline{M}) = (1 + (m-1)\Sigma_B\Sigma^{-1})\Sigma/mn \qquad (6.14)$$

$$V_I(\hat{\Sigma}_{ANOVAij}|s,\underline{M}) = [(m\Sigma_{Bii} + \Sigma_{Wii})(m\Sigma_{Bjj} + \Sigma_{Wjj}) + (m\Sigma_{Bij} + \Sigma_{Wij})^2]/m^2(n-1) + (m-1)(\Sigma_{Wii}\Sigma_{Wjj} + \Sigma_{Wij}^2)/m^2n \qquad (6.15)$$

$$\to (\Sigma_{ii}\Sigma_{jj} + \Sigma_{ij}^2 + (m-1)(\Sigma_{Bii}\Sigma_{Bjj} + \Sigma_{Bij}^2))/mn \text{ as } n \to \infty$$

Proof : Since $L_m'L_m$, $L_B'L_B$ and $L_W'L_W$ are idempotent, Cochran's theorem applies (e.g. Anderson, 1958, 7.4) and

$$n(m-1)S_s^W = \tilde{e}'L_W'L_W\,\tilde{e} \sim W_p(tr(L_W'L_W),\Sigma_W)$$

$$= W_p(n(m-1),\ \Sigma_W) \qquad (6.16)$$

$$(n-1)\ S_s^B \sim W_p(tr(P_{Bn})\ ,\ m\Sigma_B + \Sigma_W)$$

$$= W_p(n-1,\ m\Sigma_B + \Sigma_W) \qquad (6.17)$$

S_s^W and S_s^B are independent.

Hence as in Lemma 2.9

$$V_I(\hat{\Sigma}_{ANOVAij}|s,\underline{M}) = V_I(S_{sij}^B|s,\underline{M})/m^2 + (m-1)^2\ V_I(S_{sij}^W|s,\underline{M})/m^2$$

$$= ((m\Sigma_B + \Sigma_W)_{ii}(m\Sigma_B + \Sigma_W)_{jj} + (m\Sigma_B + \Sigma_W)_{ij}^2)/m^2(n-1)$$

$$+ (m-1)\ (\Sigma_{Wii}\ \Sigma_{Wjj} + \Sigma_{Wij}^2)/m^2n$$

as required.

Also

$$V_I(\hat{\mu}_{ANOVA}|s,\underline{M}) = E_I(\Sigma_W/mn|s,\underline{M}) + V_I(\Sigma\mu_i/n|s,\underline{M})$$

$$= \Sigma_W/mn + \Sigma_B/n$$

$$= (1 + (m-1)\ \Sigma_B\Sigma^{-1})\Sigma/mn$$

Note that $\hat{\mu}_{ANOVA}$ is identical to the standard estimator T_{Ym} of (5.26). The elements of Σ_{ANOVA} differ slightly from T_{Yv} and T_{XYc} defined in (5.40) and Section 5.5, since

$$T_{Yv} = (n-1)\, S^{B}_{sii} /(nm-1) + n(m-1)S^{W}_{sii}/(nm-1)$$

$$T_{XYc} = (n-1)\, S^{B}_{sij} /(nm-1) + n(m-1)\, S^{W}_{sij}/(nm-1)$$

Note that the two estimators converge as $n \to \infty$ and that from Lemma 6.2 the 'misspecification effect' of $\hat{\Sigma}_{ANOVA}$ has the familiar $1 + (m-1)\tau$ form as $n \to \infty$.

ML Estimation

Lemma 6.3

If (6.13) holds the maximum likelihood estimators of μ and Σ are

$$\hat{\mu}_{ML} = \bar{y}_s$$

$$\hat{\Sigma}_{ML} = \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij}-\bar{y}_s)(y_{ij}-\bar{y}_s)'/mn$$

Proof : This follows by generalising standard results for the case p=1 (e.g. Arnold, 1981 p.251). Note that $\hat{\Sigma}_{ML}$ is always non-negative definite and so always lies in the admissible parameter space (unlike the ML estimators of Σ_B which has a positive probability of lying on the boundary of the parameter space).

Note that $\hat{\mu}_{ML}$ is the same estimator as $\hat{\mu}_{ANOVA}$ and as in Chapter 5 and that $\hat{\Sigma}_{ML}$ is a multiple (mn-1)/mn of the standard estimators T_{Yv} and T_{XYc} in Chapter 5.

Restricted Maximum Likelihood (REML) Estimation

A number of authors (e.g. Patterson and Thompson, 1971) have argued that in estimating Σ_B and Σ_W it is better to maximise the marginal likelihood of S^{B}_{s} and S^{W}_{s} (sufficient statistics for Σ_B and Σ_W)

than to maximise the full likelihood function.

Lemma 6.4

If (6.13) holds the REML estimate of Σ is

$$\hat{\Sigma}_{REML} = \hat{\Sigma}_{ANOVA}$$

Proof : From (6.16) and (6.17) $n(m-1)\ S_s^W$ and $(n-1)\ S_s^B$ are independent Wishart random matrices and so their log joint density (marginal likelihood) is

$$-\tfrac{1}{2}\left[n(m-1)\ \log|\Sigma_W| + (n-1)\ \log\ |m\Sigma_B + \Sigma_W| + n(m-1)\ tr(S_s^W\Sigma_W^{-1})\right.$$

$$\left. + (n-1)\ tr(S_s^B\ (m\Sigma_B + \Sigma_W)^{-1})\right]$$

Hence the REML estimates of Σ_W is S_s^W and of $m\Sigma_B + \Sigma_W$ is S_s^B and hence of Σ is $S_s^B/m + (m-1)\ S_s^W/m = \hat{\Sigma}_{ANOVA}$.

Minimum Variance Unbiased Estimation

Lemma 6.5

$\bar{y}_s$ is the uniformly minimum variance linear unbiased estimate of μ. $\hat{\Sigma}_{ANOVA}$ is the uniformly minimum variance quadratic unbiased estimator of Σ. If (6.13) holds then $\bar{y}_s$ and $\hat{\Sigma}_{ANOVA}$ are the uniformly minimum variance unbiased estimators of μ and Σ respectively.

Proof : $\bar{y}_s$ is a BLUE of μ by the Gauss-Markov theorem. Tan (1978) shows that $\hat{\Sigma}_{ANOVA}$ is a BQUE of Σ. Under normality $\bar{y}_s$ and $\hat{\Sigma}_{ANOVA}$ are uniformly BUE's from the Lehman-Scheffé Theorem because $(\bar{y}_s, S_s^B, S_W^B)$ is a complete sufficient statistic for $(\mu, \Sigma_B, \Sigma_W)$ (see Arnold, 1981, p.248 for p=1) and because $\bar{y}_s$ and $\hat{\Sigma}_{ANOVA}$ are unbiased for μ and Σ.

Case 2 : m_i unequal, $\Sigma_i = \Sigma_W$

This is the conventional unbalanced one-way classification multivariate random effects model (e.g. Searle, 1956). The basic

parameters are again μ, Σ_W and Σ_B.

ANOVA Estimation

Lemma 6.6

The ANOVA estimators are

$$\hat{\mu}_{ANOVA} = \bar{y}_s$$

$$\hat{\Sigma}_{ANOVA} = (n-1)(S_s^B - S_s^W)/(m_o - m^*) + S_s^W$$

where

$$\bar{y}_s = \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij}/m_o \qquad (6.18)$$

$$S_s^B = \sum_{i=1}^{n} m_i(\bar{y}_i - \bar{y}_s)(\bar{y}_i - \bar{y}_s)'/(n-1) \qquad (6.19)$$

$$S_s^W = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'/(m_o - n) \qquad (6.20)$$

$$\bar{y}_i = \sum_{j=1}^{m_i} y_{ij}/m_i$$

$$\bar{m} = \sum m_i/n \quad , \quad m_o = n\bar{m} \quad , \quad m^* = \sum m_i^2/m_o$$

$\hat{\mu}_{ANOVA}$ and $\hat{\Sigma}_{ANOVA}$ are unbiased for μ and Σ respectively given s and $\underline{m}$.

Proof : These estimators are given by Searle (1956) and are obtained as in Lemma 6.1.

Note that $\hat{\mu}_{ANOVA}$ is again the standard estimator of Chapter 5 but that $\hat{\Sigma}_{ANOVA}$ again differs from T_{Yv} and T_{XYc}.

$$T_{Yv} = (n-1)S_{sii}^B/(m_o-1) + (m_o-n)\, S_{sii}^W/(m_o-1)$$

$$T_{XYc} = (n-1)S_{sij}^B/(m_o-1) + (m_o-n)\, S_{sij}^W/(m_o-1)$$

Note again that both estimators converge as $n \to \infty$ provided $\bar{m}$ and m^* converge.

The variance of $\hat{\mu}_{ANOVA}$ under Model I is given in Chapter 5. The variance of $\hat{\Sigma}_{ANOVA}$ is given by Searle (1956) under the assumption of normality but is not reproduced here since it is rather complicated.

ML Estimator

ML estimation is much more difficult in the unbalanced case. We initially obtain $\tilde{\mu}$ in terms of the MLE's of Σ_B and Σ_W essentially using generalised least squares (Harville, 1977).

Lemma 6.7

If $Y_{ij}|\theta_i \sim N_p(\mu_i, \Sigma_W)$, $\mu_i|M_i \sim N_p(\mu_i, \Sigma_B)$ then the maximum likelihood estimator of μ is

$$\hat{\mu}_{ML} = \left(\sum_{i=1}^{n} \hat{A}_i\right)^{-1} \sum_{i=1}^{n} \hat{A}_i \bar{y}_i$$

where $\hat{A}_i = (\hat{\Sigma}_{BML} + \hat{\Sigma}_{WML}/m_i)^{-1}$

and $\hat{\Sigma}_{BML}$ and $\hat{\Sigma}_{WML}$ are the MLE's of Σ_B and Σ_W.

Proof : Let y_i be the $pm_i \times 1$ vector, $y_i' = (y_{i1} \ldots y_{im_i}')'$.

Then $$E_I(y_i|M_i) = 1_{m_i} \otimes \mu$$

$$\Sigma_{yi} = V_I(y_i|M_i) = 1_{m_i} 1_{m_i}' \otimes \Sigma_B + I_{m_i} \otimes \Sigma_W$$

$$= P_{Wm_i} \otimes \Sigma_W + P_{Bm_i} \otimes (m_i \Sigma_B + \Sigma_W)$$

where $P_{Bm.} = 1_m 1_{m.}'/m$, $P_{Wm} = I_m - P_{Bm.}$

The log likelihood is

$$\ell = -\tfrac{1}{2}\Sigma m_i p \log 2\Pi - \tfrac{1}{2}\Sigma \log|\Sigma_{yi}| - \tfrac{1}{2}(y_i - 1_{m_i} \otimes \mu)' \Sigma_{yi}^{-1} (y_i - 1_{m_i} \otimes \mu)$$

Now $\Sigma_{yi}^{-1} = P_{Wm_i} \otimes \Sigma_W^{-1} + P_{Bm_i} \otimes (m_i \Sigma_B + \Sigma_W)^{-1}$

and $|\Sigma_{yi}| = |\Sigma^W|^{m_i - 1} |m_i \Sigma_B + \Sigma_W|$

Further $(y_i - 1_{m_i} \otimes \mu)' P_{Wm_i} \otimes \Sigma_W^{-1} (y_i - 1_{m_i} \otimes \mu)$

$$= y_i' P_{Wm_i} \otimes \Sigma_W^{-1} y_i$$

$$= \mathrm{tr}\Big[\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)' \Sigma_W^{-1}\Big]$$

and $(y_i - 1_{m_i} \otimes \mu)' P_{Bm_i} \otimes (m_i \Sigma_B + \Sigma_W)^{-1} (y_i - 1_{m_i} \otimes \mu)$

$$= \mathrm{tr}\big[m_i (\bar{y}_i - \mu)(\bar{y}_i - \mu)' (m_i \Sigma_B + \Sigma_W)^{-1}\big]$$

Hence

$$\ell = -\tfrac{1}{2} m_o p \log(2\Pi) - \tfrac{1}{2}(m_o - n) \log|\Sigma_W| - \tfrac{1}{2} \sum_i \log|m_i \Sigma_B + \Sigma_W|$$

$$- \tfrac{1}{2} \sum_i \mathrm{tr}\Big(\sum_j (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)' \Sigma_W^{-1}\Big)$$

$$- \tfrac{1}{2} \sum_i \mathrm{tr}\big[m_i (\bar{y}_i - \mu)(\bar{y}_i - \mu)' (m_i \Sigma_B + \Sigma_W)^{-1}\big]$$

$$= -\tfrac{1}{2} m_o p \log(2\Pi) - \tfrac{1}{2}(m_o - n) \log|\Sigma_W| - \tfrac{1}{2} \sum_i \log|m_i \Sigma_B + \Sigma_W|$$

$$- \tfrac{1}{2} (m_o - n) \mathrm{tr}(S_s^W \Sigma_W^{-1})$$

$$- \tfrac{1}{2} \Sigma (\bar{y}_i - \tilde{\mu})' (\Sigma_B + \Sigma_W / m_i)^{-1} (\bar{y}_i - \tilde{\mu})$$

$$- \tfrac{1}{2} (\mu - \tilde{\mu})' \Sigma_i (\Sigma_B + \Sigma_W / m_i)^{-1} (\mu - \tilde{\mu})$$

where $\tilde{\mu} = \Big[\sum_{i=1}^{n} (\Sigma_B + \Sigma_W / m_i)^{-1}\Big]^{-1} \sum_{i=1}^{n} (\Sigma_B + \Sigma_W / m_i)^{-1} \bar{y}_i$

Hence the likelihood is maximised when $\mu = \tilde{\mu}$ with $\hat{\Sigma}_{BML}$ and $\hat{\Sigma}_{WML}$ substituted for Σ_B and Σ_W i.e. $\mu = \hat{\tilde{\mu}}_{ML}$.

Note that $\hat{\mu}_{ML}$ is the linear combination of the $\bar{y}_i$ weighted according to their inverse covariance matrices, A_i , and that $\hat{\mu}_{ML} = \hat{\mu}_{ANOVA}$ only if $\hat{\Sigma}^B_{ML} = 0$. Note also that $\hat{\mu}_{ML}$ may be written as

$$\hat{\mu}_{ML} = (\Sigma\ \hat{\Lambda}_i)^{-1}\ \Sigma\ \hat{\Lambda}_i \bar{y}_i \qquad (6.21)$$

where $\hat{\Lambda}_i = \hat{\Sigma}_{BML}(\hat{\Sigma}_{BML} + \hat{\Sigma}_{WML}/m_i)^{-1}$

is the multivariate generalisation of the λ_i in Scott and Smith (1969).

Unfortunately no closed-form expressions for $\hat{\Sigma}_{BML}$ and $\hat{\Sigma}_{WML}$ are available (see e.g. Searle, 1971, p.462 for the univariate case). Instead many numerical approaches have been proposed (e.g. Hartley and Rao, 1967; Hemmerle and Hartley, 1973; Harville, 1977). We only present one iterative approach based on the EM algorithm (e.g. Demster et al, 1977, 1981).

The EM approach distinguishes between the 'incomplete data' which in our case is $\underline{y}_s = (y_{11} \ldots y_{1m_1} \ldots y_{nm_n})$ and the 'complete data' which we take to be $(\underline{y}_s, \mu_1 \ldots \mu_n)$. We make the normality assumptions of Lemma 6.7 and so the distribution of the complete (or incomplete) data is indexed by $\theta = (\mu, \Sigma_B, \Sigma_W)$. A sufficient statistic for θ, were the complete data to be observed, is $T = (\bar{\mu}, S_\mu, S_w)$ where:

$$\bar{\mu} = \sum_1^n \mu_i/n\ ,$$

$$S_\mu = \Sigma(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'/n\ ,$$

$$S_w = \Sigma\Sigma(y_{ij} - \mu_i)(y_{ij} - \mu_i)'/m_o.$$

The EM algorithm proceeds as follows:

1. Select inital values $\theta^{(0)}$ for θ, e.g. the ANOVA estimators.
   For k = 1,2...
2. E-step. Compute $T^{(k)} = E(T|\underline{y}_s, \theta = \theta^{(k-1)})$.
3. M-step. Let $\theta^{(k)}$ be the value of θ which maximises the likelihood of the complete data based on $T^{(k)}$. Repeat 2 and 3 until convergence.

This algorithm always converges to a limit $\theta^{(\infty)}$ although not necessarily to the global maximum of the (incomplete) likelihood, $\hat{\theta}_{ML}$. The convergence can, however, be slow and Thompson (1977) has pointed out that, in the

univariate version of the above problem the EM algorithm will converge slowly if Σ_B is relatively small (which is likely to be the case in cluster surveys). Clearly further numerical investigation is necessary to assess the practicality of the above approach.

We now give explicitexpressions for the E and M-steps of the algorithm.

Lemma 6.8

With the assumptions and notation above, the E-step is obtained by:

$$\bar{\mu}^{(k)} = E(\bar{\mu}. \,|\, \underline{y}_s,\ \theta = \theta^{(k-1)}) = \sum_{i=1}^{n} \hat{\mu}_i^{(k)}/n$$

$$S_\mu^{(k)} = E(S_\mu \,|\, \underline{y}_s,\ \theta = \theta^{(k-1)}) = \sum_{i=1}^{n} (\hat{\mu}_i^{(k)} - \bar{\mu}^{(k)})(\hat{\mu}_i^{(k)} - \bar{\mu}^{(k)})/n$$

$$+ (n-1) \sum_{i=1}^{n} \Lambda_i^{(k-1)} \Sigma_W^{(k-1)}/m_i n$$

$$S_w^{(k)} = E(S_w \,|\, \underline{y}_s,\ \theta = \theta^{(k-1)}) = (m_o - n)\, S_s^W/m_o + \sum_{i=1}^{n} \Lambda_i^{(k-1)} \Sigma_W^{(k-1)}/m_o$$

$$+ \Sigma m_i (I - \Lambda_i^{(k-1)})(\bar{y}_i - \mu^{(k-1)})(\bar{y}_i - \mu^{(k-1)})'(I - \Lambda_i^{(k-1)})'/m_o$$

where $\hat{\mu}_i^{(k)} = E(\mu_i \,|\, \underline{y}_s,\ \theta = \theta^{(k-1)}) = \Lambda_i^{(k-1)} \bar{y}_i + (I - \Lambda_i^{(k-1)})\mu^{(k-1)}$

$$\Lambda_i^{(k)} = \Sigma_B^{(k)} (\Sigma_B^{(k)} + \Sigma_W^{(k)}/m_i)^{-1}$$

and S_s^W is defined in (6.20).

The M-step is obtain by

$$\theta^{(k)} = T^{(k)}$$

Proof : $E(\bar{\mu} \,|\, \underline{y}_s,\ \theta) = \Sigma E(\mu_i \,|\, \underline{y}_s,\ \theta)/n$

Now μ_i is independent of y_{kj} , $i \neq k$ and so

$$E(\mu_i \,|\, \underline{y}_s,\ \theta) = E(\mu_i \,|\, y_i,\ \theta)$$

where y_i is defined in the proof of Lemma 6.7.

$$y_i|\mu_i, \theta \sim N_{pm_i}[(1_{m_i} \otimes I_p)\ \mu_i,\ I_{m_i} \otimes \Sigma_W]$$

$$\mu_i|\theta \sim N_p(\mu, \Sigma_B)$$

Hence, e.g. from the Lemma on page 4 of Lindley and Smith (1972),

$$\mu_i|y_i, \theta \sim N_p(B_i b_i, B_i) \qquad (6.22)$$

where $B_i^{-1} = (1_{m_i} \otimes I_p)'\ (I_{m_i} \otimes \Sigma_W)^{-1}(1_{m_i} \otimes I_p) + \Sigma_B^{-1}$

$$= m_i\ \Sigma_W^{-1} + \Sigma_B^{-1}$$

$$b_i = (1_{m_i} \otimes I_p)'(I_{m_i} \otimes \Sigma_W)^{-1}\ y_i + \Sigma_B^{-1}\ \mu$$

$$= m_i\ \Sigma_W^{-1}\ \bar{y}_i + \Sigma_B^{-1}\ \mu$$

$$\therefore B_i b_i = (m_i\ \Sigma_W^{-1} + \Sigma_B^{-1})^{-1}\ m_i\ \Sigma_W^{-1}(\bar{y}_i - \mu) + \mu$$

$$= \Sigma_B(\Sigma_B + \Sigma_W/m_i)^{-1}\ (\bar{y}_i - \mu) + \mu$$

$$= \Lambda_i \bar{y}_i + (I-\Lambda_i)\mu$$

where $\Lambda_i = \Sigma_B(\Sigma_B + \Sigma_W/m_i)^{-1}$

$$\therefore E(\bar{\mu}|\underline{y}_s, \theta) = \Sigma(\Lambda_i \bar{y}_i + (I-\Lambda_i)\mu)/n \text{ as required}$$

$$E(S_\mu|\underline{y}_s, \theta) = E(\Sigma(\mu_i-\bar{\mu})(\mu_i-\bar{\mu})'/n|\underline{y}_s, \theta)$$

Now (conditional on $\underline{y}_s$) the μ_i are independently distibuted as in (6.22). Hence

$$E(S_\mu|\underline{y}_s, \theta) = E(\Sigma_i (n-1)\mu_i\mu_i'/n^2 - \Sigma_{i\neq j}\ \mu_i\mu_j'/n^2|\underline{y}_s, \theta)$$

$$= (n-1)\ \Sigma_i\ (B_i b_i b_i' B_i' + B_i)/n^2 + \Sigma_{i\neq j}\ B_i b_i b_j' B_j'/n^2$$

$$= (n-1)\ \Sigma B_i/n^2 + \Sigma(\hat{\mu}_i-\bar{\mu})(\hat{\mu}_i-\bar{\mu})/n$$

in obvious notation

$$= (n-1)\ \Sigma\Lambda_i\Sigma_W/m_i n + \Sigma(\hat{\mu}_i-\bar{\mu})(\hat{\mu}_i-\bar{\mu})/n$$

as required

$$E(S_w|\underline{y}_s,\ \theta) = E(\Sigma\Sigma(y_{ij}-\mu_i)(y_{ij}-\mu_i)'/m_o|\underline{y}_s,\ \theta)$$

$$= E(\Sigma\Sigma(y_{ij}-\bar{y}_i)(y_{ij}-\bar{y}_i)'/m_o$$

$$+ \Sigma m_i(\mu_i-\bar{y}_i)(\mu_i-\bar{y}_i)'/m_o|\underline{y}_s,\theta)$$

$$= (m_o-n)\ S_s^W/M_o$$

$$+ \Sigma m_i\left[(\tilde{\mu}_i-\bar{y}_i)(\hat{\mu}_i-\bar{y}_i)' + B_i\right]/m_o$$

$$= (m_o-n)\ S_s^W/m_o + \Sigma\Lambda_i\Sigma_W/m_o$$

$$+ \Sigma m_i(I-\Lambda_i)(\bar{y}_i-\mu)(\bar{y}_i-\mu)'(I-\Lambda_i)'/m_o$$

as required

Finally, the likelihood of the complete data is the product of the m_o densities of the $y_{ij}-\mu_i$ which are IID $\sim N(0,\Sigma_W)$ and the n densities of the μ_i which are IID $\sim N(\mu,\Sigma_B)$. Hence

$$\Sigma_W^{(k)} = S_W^{(k)} \quad , \quad \mu^{(k)} = \bar{\mu}^{(k)} \quad , \quad \Sigma_B^{(k)} = S_\mu^{(k)}$$

as required.

Note that the $\tilde{\mu}_i$ above are the multivariate analogues of the predictors of μ_i in Scott and Smith (1969) which shrink $\bar{y}_i$ towards μ.

Note also that

$$\mu^{(k)} = \Sigma\, \hat{\mu}_i^{(k)}/n$$

$$= \Sigma(\Lambda_i^{(k-1)}\, \bar{y}_i + (I-\Lambda_i^{(k-1)})\, \mu^{(k-1)})/n$$

$$= \mu^{(k-1)} + (\Sigma\Lambda_i^{(k-1)}/n)\left[(\Sigma\Lambda_i^{(k-1)})^{-1}(\Sigma\Lambda_i^{(k-1)}\bar{y}_i) - \mu^{(k-1)}\right]$$

so that at convergence when $\mu^{(k)} = \mu^{(k-1)}$ we have

$$\mu^{(k)} = (\Sigma\Lambda_i^{(k)})^{-1}\, \Sigma\Lambda_i^{(k)}\bar{y}_i$$

as in (6.20).

REML Estimation

One approach would be to maximise the likelihood of $(y_{11}-\bar{y}_s \ \ldots\ y_{1m_1} - \bar{y}_s \ \ldots\ y_{nm_n} - \bar{y}_s)$ (e.g. Harville, 1977, p.325). Another approach would be to assume a flat prior for μ (Dempster, et al, 1981, p.343). In neither case would there appear to be much gain in computational efficiency (e.g. Dempster et al, 1981) nor in estimation efficiency (e.g. Harville, 1977, since the number of parameters in μ is likely to be much less than the overall degrees of freedom). Note that REML is not the same as maximising the likelihood concentrated by $\hat{\mu}_{ML}$ of Lemma 6.7.

Minimum Variance Unbiased Estimation

In the case of unequal m_i there are no *uniformly* minimum variance unbiased estimators of μ and Σ. Instead we may consider estimators which are *locally* best at given points of the parameter space. For example,

$$\tilde{\mu} = (\Sigma\, \tilde{A}_i)^{-1}\, \Sigma\, \tilde{A}_i\bar{y}_i \qquad (6.23)$$

where $\tilde{A}_i = (\tilde{\Sigma}_B + \tilde{\Sigma}_W/M_i)^{-1}$ and $\tilde{\Sigma}_B$ and $\tilde{\Sigma}_W$ are fixed 'prior values' (independent of the data), is minimum variance linear unbiased for μ when $\tilde{\Sigma}_B = \Sigma_B$ and $\tilde{\Sigma}_W = \Sigma_W$. Similarly locally minimum variance quadratic unbiased estimators of Σ_B and Σ_W are given by Lamotte (1973) for the

case of norma lity with p = 1. These estimators depend on prior values $\tilde{\Sigma}_B/\tilde{\Sigma}_W$ and $\tilde{\mu}/\tilde{\Sigma}_W^{1/2}$. The problem with such estimators is, of course, the specification of the prior values. The sensitivity of such estimators to the values is greatest when there are least restrictions (e.g. unbiasedness, invariance, quadratic) on the estimators. In the ludicrously exteme case of no restrictions, the trivial minimum variance estimators of μ and Σ are the 'prior values' $\tilde{\mu}$ and $\tilde{\Sigma}$. A number of studies have been made of the sensitivity of such locally best estimators to the prior values and of the efficiency of these estimators with respect to the ANOVA estimators. For example, the simulation studies of Hess (1979) and Swallow (1981) suggest that for p = 1 the ANOVA estimator of Σ_W is always preferable and the ANOVA estimator of Σ_B is preferable if Σ_B/Σ_W is small (say < 1) as is usually the case in cluster surveys. On the basis of large sample theory for p = 1, Seely (1979) also recommends the ANOVA estimator of Σ_W and, when the intracluster correlation is low, the ANOVA estimator of Σ_B.

One way of avoiding the dependence on prior values is to substitute estimates of these values. This will, however, generally invalidate the optimal properties of the estimators. See, for example, Fuller and Battese (1973) on the properties of $\tilde{\mu}$ above with ANOVA estimators substituted for $\tilde{\Sigma}_B$ and $\tilde{\Sigma}_W$. Alternatively an iterative procedure of resubsitution of best estimates for prior values might be used. For the estimation of Σ under normality the iterated minimum variance quadratic translation-invariant unbiased estimator turns out to be identical to the REML estimator (Searle, 1979; Rao, 1979).

In summary, in spite of the vast amount of recent literature on minimum variance estimation particularly since Rao's work on MINQUE estimation in the early 1970's, there seems little advantage for our purposes in going beyond ANOVA and ML estimation.

Case 3 : m_i and Σ_i unequal

ANOVA Estimation

In this case the moments of the Y_{ij} are

$$E(Y_{ij}) = EE(Y_{ij}|\theta_i) = E(\mu_i) = \mu \tag{6.24}$$

$$\begin{aligned} \text{cov}(Y_{ij}, Y_{k\ell}) &= \text{cov}(\mu_i, \mu_k) + E(\text{cov}(Y_{ij}, Y_{k\ell}|\theta_i, \theta_k)) \\ &= \Sigma_B + E(\Sigma_i) = \Sigma_B + \Sigma_W \quad \text{if } i=k \quad j=\ell \\ &= \Sigma_B \quad \text{if } i=k \quad j\neq\ell \\ &= 0 \quad \text{if } i\neq k \end{aligned} \tag{6.25}$$

Hence the first and second moment structure is as in Case 1 and 2 and so the ANOVA estimators are as before and are unbiased.

ML Estimation

One possible distributional assumption would be that the Y_{ij} were jointly normally distributed (unconditional on the θ_i) with the mean and covariance sturcture of (6.24) and (6.25). In this case the ML estimates would be identical to those given for Cases 1 and 2.

Alternatively we might attempt to specify the within and between cluster distributions separately. We take the within cluster distribution as before as:

$$Y_{ij}|\theta_i \sim N_p(\mu_i, \Sigma_i)$$

For tractability, we take the usual conjugate prior (e.g. Dempster, 1969, p.368) for the between cluster distribution:

$$\mu_i|\Sigma_i \sim N_p(\mu, b^2 \Sigma_i)$$

$$\Omega_i = \Sigma_i^{-1} \sim W_p(C, \Omega^{-1})$$

where (μ, b^2, c, Ω) are unknown parameters. Hence

$$\Sigma = V(Y_{ij}) = V(\mu_i) + E(\Sigma_i)$$

$$= E\ V(\mu_i|\Sigma_i) + V\ E(\mu_i|\Sigma_i) + E(\Sigma_i)$$

$$= E\left[(1 + b^2)\ \Sigma_i\right]$$

$$= (1 + b^2)\ \Omega/(c-p-1)$$

since Σ_i has an inverse Wishart distribution (e.g. Johnson and Kotz, 1974, p.164). The above distribution is rather unsatisfactory since it assumes that the between-cluster covariance structure is proportional to the within-cluster covariance structure. However, it seems the most obvious starting point.

ML estimates of μ and Σ may be obtained as in Case 2 using the EM algorithm where the complete data is now $(\underline{y}_s\ \mu_1 \ldots\ \mu_n\ \Omega_1\ \ldots \Omega_n)$. Were the complete data to be observed, the likelihood would be of the exponential family form (Dempster et al, 1977) with sufficient statistic $T = (\Sigma\Omega_i,\ \Sigma\log|\Omega_i|,\ \Sigma\mu_i'\Omega_i\mu_i,\ \Sigma\Omega_i\mu_i)$ for $\Theta = (\mu,\ b^2,\ c,\ \Omega)$(see proof of Lemma 6.9). The EM algorithm proceeds as in Case 2. Expressions for the separate steps are given in the following lemma.

Lemma 6.9

With the assumptions and notation above the E-step is obtained by

$$E(\Sigma\Omega_i|\underline{y}_s,\ \theta) = \sum_{i=1}^{n} (m_i + c)\ \Omega_i^*$$

$$E(\Sigma\log|\Omega_i|\ |\underline{y}_s,\ \theta) = np\ \log 2 + \sum_{i=1}^{n}\ \sum_{k=1}^{p} \Psi[(m_i+c+1-k)/2] + \sum_{i=1}^{n} \log|\Omega_i^*|$$

$$E(\Sigma\mu_i'\Omega_i\mu_i|\underline{y}_s,\ \theta) = \sum_{i=1}^{n} (m_i+c)\mu_i^{*\prime}\Omega_i^*\mu_i^* + pb^2 \sum_{i=1}^{n} (1+M_i b^2)^{-1}$$

$$E(\Sigma\Omega_i\mu_i|\underline{y}_s,\ \theta) = \sum_{i=1}^{n} (m_i+c)\ \Omega_i^*\mu_i^*$$

where $\mu_i^* = (\mu/m_i + b^2\bar{y}_i)/(1/m_i + b^2)$

$$\Omega_i^* = \left[\Omega + \sum_j (y_{ij}-\bar{y}_i)(y_{ij}-\bar{y}_i)' + (\bar{y}_i-\mu)(\bar{y}_i-\mu)'(1/m_i+b^2)^{-1}\right]^{-1}$$

and $\underline{\Psi}$ is the digamma function (e.g. Abramowitz and Stegun, 1964, p.258).

At the M-step the ML estimators of θ given T are

$$\hat{\mu} = (\Sigma\Omega_i)^{-1} \Sigma\Omega_i\mu_i \tag{6.26}$$

$$\hat{b}^2 = \Sigma(\mu_i-\hat{\mu})' \Omega_i(\mu_i-\hat{\mu})/np \tag{6.27}$$

$$\hat{c} = g^{-1} \left[\log |\Sigma\Omega_i/n| - (\Sigma\log|\Omega_i|)/n\right] \tag{6.28}$$

$$\hat{\Omega} = (\Sigma\Omega_i/n\hat{c})^{-1} \tag{6.29}$$

where g is the monotonic decreasing function

$$g(c) = p\log(c/2) - \sum_{j=1}^{p} \underline{\Psi} \left[(c+1-j)/2\right]$$

<u>Proof</u> : Given the assumptions above the posterior distributions of μ_i and Ω_i are (e.g. Dempster, 1969, p.369)

$$\mu_i|\Omega_i, \underline{y}_s, \theta \sim N_p(\mu_i^*, b^2 \Omega_i^{-1}/(1+m_ib^2))$$

$$\Omega_i|\underline{y}_s, \theta \sim W_p(m_i+c, \Omega_i^*)$$

Hence $E(\Sigma\Omega_i|\underline{y}_s, \theta) = \Sigma(m_i+c)\Omega_i^*$

Also from Chen (1979, Theorem 2.1)

$$E(\Sigma\log|\Omega_i||\underline{y}_s, \theta) = np\ \log 2 + \sum_i \sum_{k=1}^{p} \underline{\Psi} \left[(m_i+c+1-k)/2\right] + \Sigma\log|\Omega_i^*|$$

$$\text{Now } E(\Sigma\mu_i'\Omega_i\mu_i|\underline{y}_s, \theta) = \sum_{i=1}^{n} E\left[E(tr(\mu_i\mu_i'\Omega_i)|\Omega_i, \underline{y}_s, \theta)|\underline{y}_s, \theta\right]$$

$$= \sum_{i=1}^{n} E\left[\mu_i^{*\prime}\Omega_i\mu_i^* + tr\left[b^2\Omega_i^{-1}\Omega_i/(1+m_ib^2)\right] |\underline{y}_s, \theta\right]$$

$$= \sum_{i=1}^{n} (m_i+c)\mu_i^{*\prime}\Omega_i^*\mu_i^* + pb^2 \sum_{i=1}^{n} (1+m_ib^2)^{-1}$$

$$E(\Sigma\Omega_i\mu_i|\underline{y}_s,\ \theta) = \sum_{i=1}^{n} E\left[E(\Omega_i\mu_i|\Omega_i,\ \underline{y}_s,\ \theta)|\underline{y}_s,\ \theta\right]$$

$$= \sum_{i=1}^{n} E(\Omega_i\mu_i^*|\underline{y}_s,\ \theta)$$

$$= \sum_{i=1}^{n} (m_i+c)\Omega_i^*\mu_i^*$$

Now the joint p.d.f. of μ_i and Ω_i is given by

$$p(\mu_i|\Omega_i,\ \theta) = k_1\ b^{-p}|\Omega_i|^{\frac{1}{2}}\ \exp\left[-(\mu_i-\mu)'\Omega_i(\mu_i-\mu)/2b^2\right]$$

$$p(\Omega_i|\theta) = k_2 2^{-\frac{1}{2}cp}\ \left[\prod_{j=1}^{p}\Gamma\left[\tfrac{1}{2}(c+1-j)\right]\right]^{-1}\ |\Omega|^{\frac{1}{2}c}|\Omega_i|^{\frac{1}{2}(c-p-1)}\exp-\tfrac{1}{2}tr(\Omega\Omega_i)$$

Hence

$$p(\mu_i,\Omega_i|\theta) = k_3 2^{-\frac{1}{2}cp}\ b^{-p}\left[\prod_{j=1}^{p}\Gamma\ \left[\tfrac{1}{2}(c+1-j)\right]\right]^{-1}|\Omega|^{\ \frac{1}{2}c}|\Omega_i|^{\frac{1}{2}(c-p)}$$

$$\exp\ -\tfrac{1}{2}\left[(\mu_i-\mu)'\Omega_i(\mu_i-\mu)/b^2 + tr(\Omega\Omega_i)\right]$$

Now the likelihood of the complete data may be written

$$p(\underline{y}_s|\mu_1\ldots\mu_n\ \Omega_1\ldots\Omega_n)\ \prod_{i=1}^{n}\ p(\mu_i,\ \Omega_i|\theta)$$

where the first term does not depend on θ. Hence for the M-step we need only consider the second component of the above likelihood. The log of this expression is

$$\ell(\mu,\ b^2,\ c,\ \Omega) = k_4 - \tfrac{1}{2}ncp\ \log 2 - \tfrac{1}{2}np\ \log\ b^2 - \sum_{j=1}^{p}\ n\log\Gamma\ (\tfrac{1}{2}(c+1-j))$$

$$+ \tfrac{1}{2}nc\ \log\ |\Omega|\ +\ \tfrac{1}{2}\ (c-p)\cdot\Sigma\log\ |\Omega_i|$$

$$- \tfrac{1}{2}\Sigma(\mu_i-\mu)'\Omega_i(\mu_i-\mu)/b^2 - \tfrac{1}{2}tr(\Omega\Sigma\Omega_i)$$

This is maximised when $\mu = \tilde{\mu}$ in (6.26) since

$$\Sigma(\mu_i-\mu)'\Omega_i(\mu_i-\mu) = \Sigma(\mu_i-\hat{\mu})'\Omega_i(\mu_i-\hat{\mu}) + (\hat{\mu}-\mu)'(\Sigma\Omega_i)(\hat{\mu}-\mu)$$

Also, by direct differentiation, the MLE of b^2 is $\hat{b}^2$ in (6.27). Substituting these values the concentrated log likelihood is

$$\ell(\hat{\mu}, \hat{b}^2, c, \Omega) = k_4 - \tfrac{1}{2} ncp \log 2 - \tfrac{1}{2} np \log \hat{b}^2 - n \sum_{j=1}^{n} \log\Gamma(\tfrac{1}{2}(c+1-j))$$

$$+ \tfrac{1}{2} nc \log|\Omega| - \tfrac{1}{2} (c-p) \Sigma \log|\Omega_i| - \tfrac{1}{2} np - \tfrac{1}{2} \mathrm{tr}(\Omega \Sigma \Omega_i)$$

The MLE of Ω is obtained as for the usual MLE of the covariance matrix of an IID normal sample as in (6.29). Substituting $\hat{\Omega} = \hat{\Omega}(c) = (\Sigma \Omega_i / nc)^{-1}$ as a function of c we obtain

$$\ell(\hat{\mu}, \hat{b}^2, c, \hat{\Omega}(c)) = k_5 - \tfrac{1}{2} ncp \log 2 - n \sum_{j=1}^{p} \log \Gamma(\tfrac{1}{2}(c+1-j))$$

$$- \tfrac{1}{2} nc \log|\Sigma \Omega_i / n| + \tfrac{1}{2} ncp \log c - \tfrac{1}{2}(c-p) \Sigma \log|\Omega_i|$$

$$- \tfrac{1}{2} ncp$$

Differentiating with respect to c we obtain

$$\frac{d}{dc} \ell(\hat{\mu}, \hat{b}^2, c, \hat{\Omega}(c)) = -\tfrac{1}{2} n \sum_{j=1}^{p} \Psi(\tfrac{1}{2}(c+1-j)) - \tfrac{1}{2} n \log|\Sigma \Omega_i / n|$$

$$+ \tfrac{1}{2} np \log(c/2) + \tfrac{1}{2} \Sigma \log|\Omega_i|$$

$$= \tfrac{1}{2} n \left[g(c) - \log|\Sigma \Omega_i / n| + (\Sigma \log|\Omega_i|)/n \right]$$

Hence $\hat{c}$ is as in (6.28).

Chen (1979) shows that g(c) is a monotone decreasing function and shows that the above algorithm involving the solution of an equation $g(c) = k$ at each iteration is numerically feasible. Chen deals with a full Bayes analysis of the single sample problem of which our problem is essentially the multi-sample empirical Bayes analogue.

Note that μ_i^* shrinks $\bar{y}_i$ towards μ in estimating μ_i, but that unlike $\tilde{\mu}_i$ in Lemma 6.8 which shrinks $\bar{y}_i$ in a multivariate manner, $\bar{y}_i$ is only shrunken variate by variate (in the same proportion). This is because we have assumed that the between-cluster covariance structure

is proportional to the within-cluster covariance structure. Note also that the implied estimate of Σ^i, $\Omega_i^{*-1}/(m_i+c)$ is a weighted combination of the three estimators Ω/c, $\sum_j (y_{ij}-\bar{y}_i)(y_{ij}-\bar{y}_i)'/m_i$ and $(\bar{y}_i-\mu)(\bar{y}_i-\mu)'/b^2$. Hence the estimators 'borrow information' (Scott and Smith, 1969) not only in estimating μ but also in estimating Σ. A similar approach was taken by Novick and Jackson (1974, p.318) for the univariate case p = 1. They assumed that $\log(\Sigma_i) \sim N(\mu_\Sigma, \sigma_\Sigma^2)$ and obtained the 'regressed estimate' of Σ_i as

$$[\hat{\sigma}_\Sigma^2 \sum_j (y_{ij}-\bar{y}_i)^2/m_i + 2(m_i-1)^{-1} \hat{\mu}_\Sigma]/[\hat{\sigma}_\Sigma^2 + 2(m_i-1)^{-1}]$$

Note that there do not appear to be closed form expressions for the MLE's in the balanced case $m_i = m$ as in Case 1. For example, $\hat{\mu} = \bar{y}$ was a fixed point of the algorithm in Lemma 6.8 when $m_i = m$ whereas it is not in Lemma 6.9.

We noted above that the between-cluster distribution of (μ_i, Σ_i) was rather restrictive. We suspect in fact that it is only restrictive for estimating μ and not for Σ. We do not consider any extension to a more general distribution. We do note, however, that the non-parametric maximum likelihood approach of Laird (1978) might be useful, where we specify $y_{ij}|\theta_i \sim N_p(\mu_i, \Sigma_i)$ and where the marginal distribution of (μ_i, Σ_i) is estimated non-parametrically.

Minimum Variance Unbiased Estimation

We may write our model as in (6.12) in multivariate mixed model form (Tan, 1979) :

$$y = X\alpha + Zb + \varepsilon \qquad (6.30)$$

where $y = (y_1' \ldots y_n')'$, $y_i = (y_{i1} \ldots y_{im_i})'$, $X = 1_{m_o}$, $\alpha = \mu'$, $Z = \text{diag}_n (1_{m_o})$, $b = (b_1' \ldots b_n')'$, $b_i = \mu_i'-\mu'$, $\varepsilon = (\varepsilon_1' \ldots \varepsilon_n')'$, $\varepsilon_i = (\varepsilon_{i1} \ldots \varepsilon_{im_i})'$, $\varepsilon_{ij} = y_{ij}-\mu_i$. The covariance structure of y (given in (6.24) and (6.25)) is the same as in Cases 1 and 2 and so the minimum variance linear unbiased estimator of μ will be the same as in those Cases and is given by the Gauss-Markov Theorem applied to

the above model. This is somewhat surprising since we might expect a better estimator would be obtained by substituting $\tilde{\Sigma}_i$ for $\tilde{\Sigma}_W$ in $\tilde{A}_i$ in (6.23) (c.f. Scott and Smith, 1969). One wonders whether some kind of conditioning argument is necessary. The MINQUES of Σ_B and Σ_W (and hence Σ) will also be the same as in Cases 1 and 2 because the first and second moment structure is the same. MINQUE theory (e.g. Rao, 1971a) applies because the rows of b and ϵ are uncorrelated. MINQUES are the same as minimum variance quadratic invariant unbiased estimators (MIVQUES) if the rows of b and ϵ are independently normally distributed (Rao, 1971b). In our case, however the rows of ϵ are generally not independent and so the MIVQUE's of Σ_B, Σ_W and Σ will not be the same as in Cases 1 and 2.

An intuitive way of obtaining MIVQUE's is to use the 'dispersion-mean correspondence' of Pukelsheim (1976, 1977) (see also the 'derived model' of Brown, 1978). Any statistic which is invariant with respect to the location translation $y \to y + Xa$ may be written as a function of My (the maximal invariant) where $M = I - X(X'X)^{-1} X'$ (Pukelsheim, 1976). Since $MX = 0$ we have

$$My = MZb + M\epsilon$$

$$= U\xi \text{ , say,}$$

where $U = (MZ\ M)$, $\xi' = (b'\epsilon')$.

Now $E(My \otimes My) = E(U\xi \otimes U\xi)$

$$= (U \otimes U)E\ (\xi \otimes \xi)$$

In the univariate case we can write (Pukelsheim, 1976)

$$E(\xi \otimes \xi) = A\alpha^*$$

$$\alpha^* = (\Sigma_B \Sigma_W)'$$

Hence $E(My \otimes My) = X^*\ \alpha^*$

where $X^* = (U \otimes U)A$ (6.31)

According to the mean-dispersion correspondence we can view α^* either as the dispersion parameter of the original model (6.30) or as the mean parameter of the derived model (6.31). Since the class of quadratic invariant unbiased estimators in the original model is identical to the class of linear unbiased estimators in the dervied model the MIVQUE can be obtained by applying the Gauss-Markov Theorem to the derived model. The same result applies for the multivariate case if y is vectored beforehand. In the classical case of homogeneous variances $\Sigma_i = \Sigma_W$ (where the rows of ξ are independent) we can write the derived model as

$$My \otimes My = X^* \alpha^* + \varepsilon^*$$

where ε^* has the covariance structure given e.g. by Brown (1978, Lemma 1). When the Σ_i are unequal we may introduce random effects into the dervied model

$$My \otimes My = X^* \alpha^* + Z^* b^* + \varepsilon^*$$

where b^* is now a function of the $\Sigma_i - \Sigma_W$ just as b was a function of the $\mu_i - \mu$. MIVQUE's may then be obtained by the Gauss-Markov Theorem. We do not intend to develop the algebra here but we conjecture that the ANOVA estimators are again MIVQUE in the balanced case $m_i = m$ (See Section 6.3).

6.3 Model-Based Prediction

In this section we consider the minimum variance unbiased (MVU) (Definition 3.2) prediction of $\bar{y}$ and S (see (6.3) and (6.4)) under Model I of Section 5.1. As in Section 6.2 we assume that Assumption A (of Section 5.1) holds and we evaluate moments of predictors conditional on s and $\underline{M}$.

In Section 3.3 we used two approaches to MVU prediction: (i) we used the Lehmann-Scheffe type argument of Lemma 3.7, (ii) we obtained the MVU predictors amongst a restricted class of linear or quadratic predictors. In our present set-up we only use the first approach under the very restrictive conditions of the following Lemma.

Lemma 6.10

If Assumption A holds, $m_i = M_i = M$ $(i=1 \ldots N)$ and

$$Y_{ij}|\mu_i \sim N_p(\mu_i, \Sigma_W) \qquad i=1 \ldots N,\ j=1 \ldots M$$

$$\mu_i \sim N_p(\mu, \Sigma_B)$$

then the minimum variance unbiased predictors of $\bar{y}$ and S are

$$\hat{\bar{y}} = \bar{y}_s$$

$$\hat{S} = \left[(N-1)S_s^B + N(M-1)S_s^W\right]/(NM-1) \qquad (6.32)$$

where $\bar{y}_s$, S_s^B and S_s^W are defined in (6.7), (6.9) and (6.10).

Proof

As in Lemma 6.5 a complete sufficient statistic for $\theta = (\mu, \Sigma_W, \Sigma_B)$ is $A = (\bar{y}_s, S_s^B, S_s^W)$. In the notation of Lemma 3.7, $Y = (y_{11}' \ldots y_{1M}' \ldots y_{nM}')'$, $Z = (y_{n+1,1}' \ldots y_{NM}')'$. The joint distribution of (Y,Z) is indexed by θ and, since Y is independent of Z, A is predictive sufficient for Z (definition 1.3). Initially let

$T = \bar{y}$ then

$$T = nM\,\bar{y}_s + \sum_{i=n+1}^{N}\sum_{j=1}^{M} y_{ij}$$

and so T is a function of A and Z.

Let $\hat{T} = \bar{y}_s$ then

$$E_I(\hat{T} - T|s, \underline{M}) = \mu - \mu = 0$$

so $\hat{T}$ is an unbiased predictor of T which is a function of A and so by Lemma 3.7. $\hat{T} = \bar{y}_s$ is minimum variance unbiased for $T = \bar{y}$ as required.

Now let T = S then

$$T = \left[(n-1)S_s^B + n(M-1)S_s^W + nM(\bar{y}_s - \bar{y})(\bar{y}_s - \bar{y})' + \sum_{i=n+1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y})(y_{ij} - \bar{y})'\right]/(NM - 1) \qquad (6.33)$$

which is a function of A and Z since $\bar{y}$ is a function of A and Z.

$$E_I(T|s, \underline{M}) = (N-1)M\,\Sigma_B/(NM - 1) + \Sigma_W$$

Also from the proof of Lemma 6.1.

$$E_I(S_s^B|s,\underline{M}) = M\,\Sigma_B + \Sigma_W$$

$$E_I(S_s^W|s,\underline{M}) = \Sigma_W$$

Hence $E_I(\hat{S}|s,\underline{M}) = \left[(N-1)(M\,\Sigma_B + \Sigma_W) + N(M-1)\,\Sigma_W\right]/(NM-1)$

$$= (N-1)M\,\Sigma_B/(NM-1) + \Sigma_W$$

Therefore $\hat{S}$ is an unbiased predictor of S, which is a function of A and so by Lemma 3.7, S is minimum variance unbiased for $\hat{S}$. Note

$\hat{\bar{y}} = \hat{\mu}_{ANOVA}$ (in (6.7)) and that $\hat{S} - \hat{\Sigma}_{ANOVA} = O_p(N^{-1})$ (see (6.8)).

The above result cannot be extended to the case of unequal m_i, because no complete sufficient statistic exists, nor to the case $m_i = m$, $M_i = M$, $m < M$, because A will not be predictive sufficient for Z. Instead we restrict our attention to linear or quadratic predictors.

Lemma 6.11

If Assumption A holds then the (locally) minimum variance linear unbiased predictor of $\bar{y}$ is

$$\hat{\bar{y}} = \left[m_o \bar{y}_s + \sum_{i=1}^{n} (M_i - m_i) \hat{\mu}_i + \sum_{i=n+1}^{N} M_i \hat{\mu} \right] / M_o$$

where $\hat{\mu}_i = \Lambda_i \bar{y}_i + (I - \Lambda_i) \hat{\mu}$

$$\hat{\mu} = (\Sigma \Lambda_i)^{-1} \Sigma \Lambda_i \bar{y}_i$$

$$\Lambda_i = \Sigma_B (\Sigma_B + \Sigma_W / m_i)^{-1}$$

Proof

As in the proof of Theorem 3.6. $\bar{y}$ is a minimum variance linear unbiased predictor iff $a'\hat{\bar{y}}$ is a minimum variance linear unbiased predictor of $a'\bar{y}$, where a is an arbitrary p×1 vector. The latter predictor may be obtained from Theorem 2.1. of Royall (1976) as

$$a'\hat{\bar{y}} = \left[m_o a' \bar{y}_s + \gamma' (X_{II} \hat{\beta} + V_{II \cdot I} V_I^{-1} (Y_I - X_I \hat{\beta}) \right] / M_o$$

where $\gamma = 1_{M_o - m_o} \otimes a$

$$Y_I = (y_1' \ldots y_n')' \quad , \quad y_i = (y_{i1}' \ldots y_{im_i}')' \quad i=1 \ldots n$$

$$Y_{II} = (y_{\bar{1}}' \ldots y_{\bar{n}}' \; y_{n+1}' \ldots y_N')' \quad , \quad y_{\bar{i}} = (y_{im_i+1}' \ldots y_{iM_i}')' \; i=1 \ldots n$$

$$y_i = (y_{i1}' \ldots y_{iM_i}')' \quad i=n+1 \ldots N$$

$$X_I = 1_{m_o} \otimes I_p \quad , \quad X_{II} = 1_{M_o - m_o} \otimes I_p$$

$$\beta = \mu$$

$$V_I = \text{var}(Y_I) \qquad V_{II \cdot I} = \text{cov}(Y_{II}, Y_I)$$

$$\hat{\beta} = (X_I' V_I^{-1} X_I)^{-1} X_I' V_I^{-1} Y_I$$

V_I and $V_{II \cdot I}$ may be obtained from (6.25) as

$$V_I = \bigoplus_{i=1}^{n} (I_{m_i} \otimes \Sigma_W + J_{m_i} \otimes \Sigma_B)$$

$$= \bigoplus_{i=1}^{n} (P_{Wm_i} \otimes \Sigma_W + P_{Bm_i} \otimes (m_i \Sigma_B + \Sigma_W))$$

$$V_{II \cdot I} = \begin{bmatrix} \bigoplus_{i=1}^{n} J_{M_i - m_i, m_i} \otimes \Sigma_B \\ \\ O_{kp}, \; m_{op} \end{bmatrix}$$

where $\oplus$ is the direct sum

$J_{m,n}$ is the $m \times n$ matrix of ones, $J_m = J_{m,m}$

$P_{Bm} = J_m/m \quad , \quad P_{Wm} = I_m - P_{Bm}$

$O_{m,n}$ is the $m \times n$ matrix of zeros

$$k = \sum_{i=n+1}^{N} M_i$$

Hence $V_I^{-1} = \bigoplus \left[P_{Wm_i} \otimes \Sigma_W^{-1} + P_{Bm_i} \otimes (m_i \Sigma_B + \Sigma_W)^{-1} \right]$

$$X_I' V_I^{-1} X_I = \sum_{i=1}^{n} m_i (m_i \Sigma_B + \Sigma_W)^{-1}$$

$$= \Sigma_B^{-1} \sum_{i=1}^{n} \Lambda_i \quad \text{assuming } \Sigma_B \text{ is non-singular}$$

$$X_I' V_I^{-1} Y_I = \sum_{i=1}^{n} m_i (m_i \Sigma_B + \Sigma_W)^{-1} \bar{y}_i = \Sigma_B^{-1} \sum_{i=1}^{n} \Lambda_i \bar{y}_i$$

Hence $\hat{\beta} = (\Sigma\Lambda_i)^{-1}(\Sigma\Lambda_i\ \bar{y}_i) = \hat{\mu}$

Now $$V_{II \cdot I}\ V_I^{-1} = \begin{bmatrix} \bigoplus_{i=1}^{n} J_{M_i - m_i, m_i} \otimes \Lambda_i / m_i \\ 0_{kp}\ , m_o p \end{bmatrix}$$

Hence $$V_{II \cdot I}\ V_I^{-1}\ (Y_I - X_I\ \hat{\beta}) = \begin{bmatrix} 1_{M_1 - m_1} \otimes \Lambda_1 (\bar{y}_1 - \hat{\mu}) \\ \vdots \\ 1_{M_n - m_n} \otimes \Lambda_n (\bar{y}_n - \hat{\mu}) \\ 0_{kp,1} \end{bmatrix}$$

$$X_{II}\ \hat{\beta} = 1_{M_o - m_o} \otimes \hat{\mu}$$

Hence $\gamma'(X_{II}\ \hat{\beta} + V_{II \cdot I}\ V_I^{-1}(Y_I - X_I\ \hat{\beta}))$

$$= (M_o - m_o)a'\hat{\mu} + \sum_{i=1}^{n} (M_i - m_i)\ a'\ \Lambda_i(\bar{y}_i - \hat{\mu})$$

Hence

$$a'\hat{\bar{y}} = a'\left[m_o\ \bar{y}_s + \sum_{i=1}^{n} (M_i - m_i)\hat{\mu}_i + \sum_{i=n+1}^{N} M_i\ \hat{\mu}\right]/M_o$$

as required.

Note that the predictor in Lemma 6.11 has a natural interpretation. The value of y_{ij} for sampled units is predicted by y_{ij}, the value of y_{ij} for non-sampled units in the i^{th} sampled cluster is predicted by $\hat{\mu}_i$, the minimum variance linear unbiased predictor of μ_i (proof omitted) and the value of y_{ij} for units is non-sampled clusters is predicted by $\hat{\mu}$, the minimum variance linear unbiased estimator of μ (see 6.23). Note, as in Section 6.2, that it seems unsatisfactory that Λ_i is not equal to $\Sigma_B(\Sigma_B + \Sigma_i/m_i)^{-1}$ which would be the case if we had conditioned on the Σ_i.

We now turn to the prediction of S. From Lemma 6.11. we see that there is only a uniformly minimum variance linear unbiased predictor of $\bar{y}$ in the case $m_i = M_i = M$ (i=1 ... N). On the basis of the similarity between the results for estimating μ and Σ in Section 6.2 we conjecture that there is also only a uniformly minimum variance

quadratic unbiased predictor of S in this case. Because the prediction of S is more difficult than that of $\bar{y}$ we restrict ourselves to this case in the following Lemma and also assume p=1 (i.e. y_{ij} is univariate). Note, however, that the conditions of Lemma 6.12 are much weaker than those in Lemma 6.10.

Lemma 6.12

If Assumption A holds, $m_i = M_i = M$ (i=1 ... N) and p=1 then $\hat{S}$ (defined in 6.3.2) is a uniformly minimum variance quadratic predictor of S.

Proof

As in (6.33) we may write

$$S = A_s + B_s$$

where

$$A_s = \left[\sum_{i=1}^{n} \sum_{j=1}^{M} y_{ij}^2 - n^2 M \bar{y}_s^2 / N \right] / (NM-1)$$

$$B_s = \left[\sum_{i=n+1}^{N} \sum_{j=1}^{M} y_{ij}^2 - 2n(N-n)\, M\, \bar{y}_s\, \bar{y}_{\bar{s}} / N - (N-n)^2 M\, \bar{y}_{\bar{s}}^2 / N \right] / (NM-1)$$

$$y_{\bar{s}} = \sum_{i=n+1}^{N} \sum_{j=1}^{M} y_{ij} / (N-n)M$$

A_s depends only on the units in the sample and so we may write any *quadratic predictor* Q of S as

$$Q = A_s + \sum_{i=1}^{n} \sum_{j=1}^{M} \sum_{k=1}^{n} \sum_{\ell=1}^{M} a_{ijk\ell}\, y_{ij}\, y_{k\ell} \qquad (6.34)$$

We assume, without loss of generality, that $a_{ijk\ell} = a_{k\ell ij}$.

If Q is unbiased for S then

$$E_I\Big(\sum_{ijk\ell} a_{ijk\ell}\, y_{ij}\, y_{k\ell} \,\Big|\, s, \underline{M} \Big) = E_I(B_s \,|\, s, \underline{M}) \qquad (6.35)$$

Evaluating both sides of (6.35)

$$E_I\Big(\sum_{ijk\ell} a_{ijk\ell}\, y_{ij}\, y_{k\ell} \,\Big|\, s, \underline{M}\Big) = \Big(\sum_{ijk\ell} a_{ijk\ell}\Big)\mu^2 + \Big(\sum_{ij\ell} a_{iji\ell}\Big)\Sigma_B + \Big(\sum_{ij} a_{ijij}\Big)\Sigma_W \tag{6.36}$$

using the first and second moment structure of y_{ij} in (6.24) and (6.25).

$$\begin{aligned} E_I(B_s \,|\, s, \underline{M}) &= \{\big[-n(N-n)M/N\big]\mu^2 + \big[(N-n)(N-1)M/N\big]\Sigma_B \\ &\quad + \big[(N-n)(NM-1)/N\big]\Sigma_W\}/(NM-1) \\ &= \lambda_1\,\mu^2 + \lambda_2\,\Sigma_B + \lambda_3\,\Sigma_W \quad \text{say} \end{aligned} \tag{6.37}$$

If Q is uniformly unbiased we may equate coefficients in (6.36) and (6.37) to obtain

$$\sum_{ijk\ell} a_{ijk\ell} = \lambda_1 \tag{6.38}$$

$$\sum_{ij\ell} a_{iji\ell} = \lambda_2 \tag{6.39}$$

$$\sum_{ij} a_{ijij} = \lambda_3 \tag{6.40}$$

Now

$$\begin{aligned} E_I\big[(Q-S)^2 \,|\, s, \underline{M}\big] &= V_I\big[(Q-A_s)\,|\, s, \underline{M}\big] + V_I\big[(S-A_s)\,|\, s, \underline{M}\big] \\ &\quad - 2\,\mathrm{cov}_I\big[Q - A_s,\, S - A_s \,|\, s, \underline{M}\big] \end{aligned} \tag{6.41}$$

Let us now consider the third and fourth moment structure of y_{ij}. As in Section 5.4. define the first four within cluster cumulants as

$$\mu_i = E_I(Y_{ij} \,|\, \theta_i)$$

$$\sigma_i^2 = E_I((Y_{ij} - \mu_i)^2 \,|\, \theta_i)$$

$$k_{3i} = E_I((y_{ij} - \mu_i)^3 \,|\, \theta_i)$$

$$k_{4i} = E_I((y_{ij} - \mu_i)^4 \,|\, \theta_i) - 3\,\sigma_i^4$$

and define the between cluster parameters

$$k_{3W} = E_I(k_{3i}) \qquad k_{4W} = E_I(k_{4i})$$

$$\gamma = V_I(\sigma_i^2) \qquad k_{3B} = E_I((\mu_i-\mu)^3)$$

$$k_{4B} = E\left[(\mu_i-\mu)^4\right] - 3\,\sigma_B^4 \qquad \text{where } \sigma_B^2 = \Sigma_B$$

$$C_1 = \mathrm{cov}_I((\mu_i-\mu)^2,\ \sigma_i^2) \qquad C_2 = \mathrm{cov}_I(\mu_i,\ k_{3i})$$

$$C_3 = \mathrm{cov}_I(\mu_i,\ \sigma_i^2)$$

These do not depend on M_i since Assumption A holds.

We may now write

$$V_I\left[(Q-A_s)\,|\,s,\ \underline{M}\right] = \nu(Q-A_s) + \Big(\sum_{ijk} a_{ijij}\,a_{ikik} + 2\sum_{ij\ell} a_{iji\ell}^2\Big)\gamma$$

$$+ \Big(2\sum_{ijk\ell} a_{ijij}\,a_{iki\ell} + 4\sum_{ijk\ell} a_{ijik}\,a_{iji\ell}\Big)C_1$$

$$+ \Big(8\sum_{ijk\ell m} a_{ijik}\,a_{ij\ell m}\Big)C_3\mu$$

$$+ 4\Big(\sum_{ijk} a_{ijij}\,a_{ijik}\Big)C_2 \qquad (6.42)$$

where $\nu(Q-A_s)$ is the variance of $Q-A_s$ obtained by setting $\sigma_i^2 = \sigma_W^2 = \Sigma_W$, $k_{3i} = k_{3W}$, $k_{4i} = k_{4W}$ in Model I. Further, subject to (6.38) - (6.40),

$$\mathrm{cov}_I\left[Q-A_s,\ S-A_s\,|\,s,\ \underline{M}\right] = \left[-2(N-n)\mu/N(NM-1)\right]\left[\lambda_3\,k_{3W} + 2\lambda_1\sigma_W^2\mu + 2(\lambda_2 + M\lambda_3)C_3\right.$$

$$\left. + M\lambda_2\,k_{3B} + 2M\lambda_1\sigma_B^2\mu\right]$$

which does not depend on the $a_{ijk\ell}$. Also $V_I((S-A_s)\,|\,s,\ \underline{M})$ does not depend on the $a_{ijk\ell}$ and so from (6.41) the minimum variance quadratic unbiased predictor of S is obtained by minimising $V_I(Q-A_s\,|\,s,\ \underline{M})$ given in (6.42). Now from theory on the standard random effects model (Graybill, 1954) the minimum variance quadratic unbiased estimator of $E_I(B_s\,|\,s,\ \underline{M})$ is

$$\hat{Q} - A_s = \lambda_1\,\hat{\mu}^2 + \lambda_2\,\hat{\sigma}_B^2 + \lambda_3\,\hat{\sigma}_W^2$$

where $\hat{\mu}^2 = \bar{y}_s^2 - S_s^B/nM$

$\hat{\sigma}_B^2 = (S_s^B - S_s^W)/M$

$\hat{\sigma}_W^2 = S_s^W$

Hence $\nu(\hat{Q} - A_s) \leq \nu(Q - A_s)$

We may minimise the remaining terms in (6.42) by using the fact that for n pairs (α_i, β_i) the minimum of $\Sigma\alpha_i\beta_i$ subject to $\Sigma\alpha_i = k_1$ and $\Sigma\beta_i = k_2$ occurs when $\alpha_i = k_1/n$, $\beta_i = k_2/n$. Hence the minimum of $V_I[(Q - A_s)|s, \underline{M}] - \nu(Q - A_s)$ subject to (6.38) - (6.40) occurs when

$$a_{ijij} = \lambda_3/nM$$

$$a_{iji\ell} = (\lambda_2 - \lambda_3)/nM(M-1) \qquad j \neq \ell$$

$$a_{ijk\ell} = (\lambda_1 - \lambda_2)/n(n-1)M^2 \qquad i \neq k$$

But in this case $Q - A_s = \hat{Q} - A_s$ and so $V_I(Q - A_s|s, \underline{M})$ is minimised subject to (6.38) - (6.40) when $Q = \hat{Q}$. The proof is completed by noting that $\hat{Q} = S$.

6.4 Design-Based Estimation

We now consider estimators of $\bar{y}$ and S (see 6.3 and 6.4) which might be considered appropriate on the basis of their properties with respect to the randomisation distribution induced by a given sampling design.

Simple Random Sampling at Both Stages

n clusters are selected by SRSWOR. Within the sampled clusters SRSWOR's of m_i units are selected independently where i is the (population) cluster label. The first-order inclusion probability of unit (ij) is

$$\pi_{(ij)} = m_i n/M_i N$$

The design-unbiased *Horvitz-Thompson estimator* (e_1 of Section 3.4) of $\bar{y}$ is

$$\hat{\bar{y}}_{HT} = \sum_{i=1}^{n} M_i \bar{y}_i / n\bar{M}$$

where $\bar{M} = M_o/N$

The *ratio-type estimator* (e_{10} of Section 3.4) of $\bar{y}$ is

$$\hat{\bar{y}}_R = \sum_{i=1}^{n} M_i \bar{y}_i / n \bar{M}_s$$

where $\bar{M}_s = \sum_{1}^{n} M_i / n$

The *expansion estimator* (Royall, 1976b) of $\bar{y}$ is

$$\hat{\bar{y}}_E = \sum_{i=1}^{n} m_i \bar{y}_i / m_o = \bar{y}_s$$

Royall (1976b) compared these three estimators of $\bar{y}$ with the model-based predictor of Lemma 6.11 in the univariate case. The expansion estimator is the MVLUE if $\Sigma^B = 0$. The Horvitz-Thompson estimator is model-biased unless $\bar{M}_s = \bar{M}$ in which case $\hat{\bar{y}}_{HT} = \hat{\bar{y}}_R$. The ratio-type estimator may be written

$$\hat{\bar{y}}_R = \left[m_o \bar{y}_s + \sum_{i=1}^{n} (M_i - m_i)\bar{y}_i + \sum_{i=n+1}^{N} M_i \left(\sum_{1}^{n} M_i \bar{y}_i / n\bar{M}_s\right)\right] / M_o$$

Comparing this with Lemma 6.11, $\hat{\bar{y}}_R$ may be viewed as the model-based predictor of $\bar{y}$ which predicts μ_i by $\bar{y}_i$ and estimates μ by the weighted mean $\Sigma M_i \bar{y}_i / \Sigma M_i$. If, however, the intracluster correlation is low the optimal weights should be closer to m_i than M_i.

In the multivariate case the design-based estimators estimate $\bar{y}$ 'variable by variable ' whereas the model-based predictors 'borrow information' between variables unless the Λ_i are diagonal e.g. if Σ_B is proportional to Σ_W.

In the case $m_i = M_i = M$ the design-based estimators $\hat{\bar{y}}_{HT}$, $\hat{\bar{y}}_R$ and $\hat{\bar{y}}_E$ are all equal to $\bar{y}_s$, the uniformly minimum variance unbiased

model-based predictor. It is straightforward to verify also that in this case the design-based estimators $e_2(S)$, $e_3(S)$ and $e_{12}(S)$ of Section 3.4 are all equal to $\hat{S}$ (see 6.32), the uniformly minimum variance unbiased model-based predictor. Note that this estimator is also proposed for this case by Cochran (1977, p. 239).

For general m_i and M_i, expressions for the estimators of S in Section 3.4 are complicated. For example, using the second-order inclusion probabilities

$$\pi_{(ij)(k\ell)} = \begin{cases} m_i n/M_i N & i=k \quad j=\ell \\ m_i(m_i-1)n/M_i(M_i-1)N & i=k \quad j\neq\ell \\ m_i m_k n(n-1)/M_i M_k N(N-1) & i\neq k \end{cases}$$

we may evaluate e_3 as

$$e_3(S) = N\Big\{ \sum_{i=1}^{n} M_i\left[m_i(M_i-1)(n-1) + (m_i-1)(N-1)(n\bar{M}_s - M_i)\right] S_{si}/(m_i-1)(n-1)$$
$$+ (N-1)n\bar{M}_s \sum_{i=1}^{n} M_i(\bar{y}_i - \hat{\bar{y}}_R)(\bar{y}_i - \hat{\bar{y}}_R)'/(n-1)\Big\}/nM_o(M_o-1)$$

where $S_{si} = \sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'/m_i$

If we let $n,N \to \infty$ and assume the M_i are bounded then

$$e_3(S) = \frac{\bar{M}_s^2}{\bar{M}^2}\left(\frac{\sum_1^n M_i S_{si}}{\sum_1^n M_i} + \frac{\sum_1^n M_i(\bar{y}_i - \hat{\bar{y}}_R)(\bar{y}_i - \hat{\bar{y}}_R)'}{\sum_1^n M_i}\right)$$

which we might view as the analogue of the Horvitz-Thompson estimator. The analogue of the ratio-type estimator is then given by $e_{12}(S)$ which for large n and N is

$$e_{12}(S) = \frac{\bar{M}^2}{\bar{M}_s^2} e_3(S)$$

The analogue of the expansion estimator is the standard estimator of S discussed in Chapter 5.

Note that the ratio-type estimators of $\bar{y}$ and S will be equal to the standard estimators of Chapter 5 (approximately equal for S) under proportionate allocation $m_i \alpha M_i$.

PPS Sampling

n clusters are selected *with replacement*. The i^{th} cluster is selected with probability M_i/M_o (at each draw). Within the sampled clusters SRSWOR's of m_i units are selected independently.

The usual design-unbiased estimator of $\bar{y}$ (which is a combination of $e_6(\bar{y})$ at the first stage and $e_1(\bar{y})$ at the second stage) is

$$\hat{\bar{y}}_{PPS} = \sum_{i=1}^{n} \bar{y}_i/n$$

This would be the natural model-based estimator of μ if Σ_B was large relative to Σ_W. Its interpretation as a model-based predictor of $\bar{y}$ is less obvious (Royall, 1976b, p. 660).

We cannot immediately apply the definitions in Section 3.4. to obtain a design-based estimator of S because this design is a combination of with and without-replacement designs. Let us now distinguish between

$$\bar{y}_i = \sum_{i=1}^{M_i} y_{ij}/M_i$$

and

$$\bar{y}_{si} = \sum_{i=1}^{m_i} y_{ij}/m_i$$

then we may write (from 6.4)

$$S = \left[\sum_{i=1}^{N} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)' + \sum_{i=1}^{N} M_i(\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})' \right] /(M_o - 1) \qquad (6.43)$$

Let us consider estimating the second term, since it is rather more difficult than the first. If we have no subsampling i.e. $m_i = M_i$ then we may estimate this term by generalising the estimators e_7 or e_8 of

Section 3.4. For

$$\sum_{i=1}^{N} M_i(\bar{y}_i - \bar{\bar{y}})(\bar{y}_i - \bar{\bar{y}})' = \sum_{1}^{N} M_i \bar{y}_i \bar{y}_i' - (\Sigma M_i\bar{y}_i)(\Sigma M_i\bar{y}_i')/M_o$$

and an unbiased estimator analogous to $e_7(S_{11})$ is

$$\sum_{i=1}^{n} M_i \bar{y}_i \bar{y}_i'/n\, p_i - \sum_{i\neq j} M_i \bar{y}_i M_j \bar{y}_j'/n(n-1)p_i\, p_j\, M_o$$

where $p_i = M_i/M_o$.

This reduces to

$$M_o \sum_{i=1}^{n} (\bar{y}_i - \bar{\bar{y}})(\bar{y}_i - \bar{\bar{y}})'/(n-1)$$

where $$\bar{\bar{y}} = \sum_{i=1}^{n} \bar{y}_i/n$$

In general, however, we have $\bar{m}_i < M_i$ and do not observe the $\bar{y}_i$. Let us instead consider

$$s_B = M_o \sum_{i=1}^{n} (\bar{y}_{si} - \hat{\bar{y}}_{pps})(\bar{y}_{si} - \hat{\bar{y}}_{pps})'/(n-1)$$

The expectation of s_B over stages I and II of the design is

$$E_p(s_B) = M_o\, E_I\, E_{II}\left[\left(\frac{n-1}{n}\right)\sum_{i=1}^{n} \bar{y}_{si} \bar{y}_{si}' - \frac{1}{n}\sum_{i\neq j} \bar{y}_{si} \bar{y}_{sj}\right]/(n-1)$$

$$= M_o E_I\left[\left(\frac{n-1}{n}\right)\sum_{i=1}^{n} (\bar{y}_i \bar{y}_i' + S_i(1-f_i)/m_i) - \frac{1}{n}\sum_{i\neq j} \bar{y}_i \bar{y}_i\right]/(n-1)$$

where $$S_i = \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'/(M_i - 1) \quad , \; f_i = m_i/M_i$$

$$\therefore \quad E_p(s_B) = M_o E_I\left[\sum_{i=1}^{n} (\bar{y}_i - \bar{\bar{y}})(\bar{y}_i - \bar{\bar{y}})' + \left(\frac{n-1}{n}\right)\Sigma\, S_i(1-f_i)/m_i\right]/(n-1)$$

$$= \sum_{i=1}^{N} M_i(\bar{y}_i - \bar{\bar{y}})(\bar{y}_i - \bar{\bar{y}})' + M_o n \sum_{i=1}^{N} S_i(1-f_i)M_i/m_i\; n\; M_o$$

$$= \sum_{i=1}^{N} M_i(\bar{y}_i - \bar{\bar{y}})(\bar{y}_i - \bar{\bar{y}})' + \sum_{i=1}^{N}\sum_{j=1}^{M_i} (M_i - m_i)(y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'$$

$$/m_i(M_i - 1) \qquad (6.44)$$

Finally we note that for constants $\alpha_1 \ldots \alpha_N$

$$E_p\Big(\sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{si})(y_{ij} - \bar{y}_{si})'\Big)$$

$$= E_I\Big(\sum_{i=1}^{n} \alpha_i (m_i - 1)\, S_i\Big)$$

$$= n \sum_{i=1}^{N} \alpha_i (m_i - 1) M_i\, S_i / M_o \qquad (6.45)$$

Hence combining (6.43) - (6.45) a design-unbiased estimator of S is

$$\hat{S}_{pps} = \Bigg[\sum_{i=1}^{n} \left(\frac{M_o(M_i - 1)}{n(m_i - 1)M_i}\right)\left(1 - \frac{M_i - m_i}{m_i(M_i - 1)}\right) \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{si})(y_{ij} - \bar{y}_{si})'$$

$$+ M_o \sum_{i=1}^{n} (\bar{y}_{si} - \hat{\bar{y}}_{pps})(\bar{y}_{si} - \hat{\bar{y}}_{pps})'/(n-1)\Bigg]/(M_o - 1)$$

$$= M_o\Bigg[\sum_{i=1}^{n} \sum_{j=1}^{\dot{m}_i} (y_{ij} - \bar{y}_{si})(y_{ij} - \bar{y}_{si})'/nm_i$$

$$+ \sum_{i=1}^{n} (\bar{y}_{si} - \hat{\bar{y}}_{pps})(\bar{y}_{si} - \hat{\bar{y}}_{pps})'/(n-1)\Bigg]/(M_o - 1)$$

Hence, as in $\hat{\bar{y}}_{pps}$, $\hat{S}_{pps}$ gives unit weights to the within cluster sample moments. Again this might be a plausible model-based estimator of Σ if the between-cluster variation is large relative to the within-cluster variation.

6.5 Conclusion

In this chapter we have considered estimators of μ and Σ and predictors of $\bar{y}$ and S which might be used in place of the standard estimators of Chapter 5.

In the case of equal cluster sizes and no second-stage subsampling we show that standard design-based estimators of $\bar{y}$ and S have optimal properties as model-based predictors under Assumption A. These

estimators also have optimal properties as estimators of μ and Σ for large N and are equal to the standard estimators of Chapter 5 for large n.

In other cases the problems of optimal estimation and prediction are less easy. Under Assumption A both problems generally involve a, possibly iterated, two-stage procedure (i) prediction of θ_i (= (μ_i, Σ_i) say) and (ii) estimation of μ and Σ or prediction of $\bar{y}$ and S. The prediction of θ_i generally involves pooling across clusters. There is clearly room for some ad hoc simplifications to bring out this structure more clearly in the estimation process. It would also be of interest to compare, probably by simulation, the efficiencies of the different estimators under Assumption A.

The design-based estimators of $\bar{y}$ and S generally differ between those based on srs at the first stage which weight within-cluster moments by M_i and those based on PPS at the first stage which weight by unity. Under Assumption A with relatively low intracluster correlation, model-based considerations suggest that it may be better to weight by m_i as in the standard estimators of Chapter 5.

Extension of the model-based approach in this Chapter to the case when Assumption A does not hold would be of most interest for the case when Assumption B also does not hold, for in this situation it appears that the standard estimators of Chapter 5 could go badly wrong. The theory would presumably be easiest in the case of single-stage cluster sampling with $m_i = M_i$ when we might summarise information on the i^{th} cluster by $(\bar{y}_i, S_i)$, say, and possibly refer to the approach of Chapters 2-4 with $x_{1i} = \text{vec}(\bar{y}_i, S_i)$ and $x_{2i} = M_i$.

CHAPTER 7. MULTIVARIATE METHODS UNDER TWO-STAGE SAMPLING

In Chapters 5 and 6 we considered the estimation of μ and Σ and the prediction of $\bar{y}$ and S. In this chapter we consider the estimation of functions of Σ, viz correlation coefficients (Section 7.1), regression coefficients (Section 7.2) and principal components (Section 7.3) and the estimation of parameters in a factor analysis model for Σ (Section 7.4). We no longer consider the prediction problem.

7.1 Correlation Coefficients

As in Section 5.5 we assume that a pair (x_{ij},y_{ij}) is associated with the j^{th} unit in the i^{th} cluster. We consider the estimation of $\rho = \sigma_{XY}/\sigma_X\sigma_Y$ where σ_{XY},σ_X^2 and σ_Y^2 are the second moments of X_{ij} and Y_{ij} about their means in the distribution f_o defined in (5.3). We shall only consider the properties of the standard estimator of ρ :

$$T_{XYr} = T_{XYc}/(T_{Xv}T_{Yv})^{\frac{1}{2}} \tag{7.1}$$

where T_{Xv} and T_{Yv} are defined as in (5.40) and T_{XYc} is defined in Section 5.5. We do not consider alternative estimators of ρ as in Chapter 6 because in the simple cases where we obtained a simple closed-form estimator of Σ the estimator was approximately equal to T_{XYc} anyway.

In this chapter we suppose that Assumption B of Section 5.2 holds. In this case it follows from Lemma 5.5 that T_{XYr} will be approximately unbiased for ρ for large n. Frankel (1971) simulated the p-distribution of T_{XYr} for various variables for self-weighting stratified (single-stage) clustered designs for a given finite population. We might conjecture from Lemma 5.11 that this p-distribution might resemble our ξ-distribution under Assumption B. Frankel (1971, p.52) found that with $n = 60$ (and two psu's per stratum, and $m_o \doteq 850$) the average bias of T_{XYr} was about 10% of the average standard error. We therefore propose to measure the effect of misspecifying Model II as Model I by the misspecification effect of Definition 5.1. As in Chapter 5 we shall approximate this misspecification effect by that of

$$T_{XYr}^{\sim} = T_{XYc}^{\sim}/(T_{Xv}^{\sim}T_{Yv}^{\sim})^{\frac{1}{2}} \tag{7.2}$$

where $T_{Xv}^{\sim}$ and $T_{Yv}^{\sim}$ are as defined in Section 5.4 and $T_{XYc}^{\sim}$ as in Section 5.5.

Lemma 7.1

If Assumption B holds

$$\text{meff}(T_{XYr}^{\sim}|s,\underline{M}) \doteq 1 + \sum_{i=1}^{n} m_i(m_i-1)\tau_{XYr}^{\sim}(M_i)/m_o \tag{7.3}$$

where

$$\tau_{XYr}^{\sim}(M_i) = \text{corr}_I\left[h_r(X_{ij},Y_{ij}),h_r(X_{ij'},Y_{ij'})\middle|M_i\right] \quad j \neq j' \tag{7.4}$$

$$h_r(X,Y) = \left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right) - \frac{1}{2}\rho\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right] \tag{7.5}$$

Proof:

This follows from Lemma 5.5 (by noting that C1 and C2 apply) and Lemma 5.9 by noting that

$$T_{XYr}^{\sim} = g(T_{XYc}^{\sim},T_{Xv}^{\sim},T_{Yv}^{\sim}) \tag{7.6}$$

where

$$g(x,y,z) = xy^{-\frac{1}{2}}z^{-\frac{1}{2}} \tag{7.7}$$

$$g_x(m_o\underline{\mu}) = g_x(\sigma_{XY},\sigma_X^2,\sigma_Y^2) = \sigma_X^{-1}\sigma_Y^{-1}$$

$$g_y(m_o\underline{\mu}) = -\frac{1}{2}\sigma_{XY}\sigma_X^{-3}\sigma_Y^{-1}$$

$$g_z(m_o\underline{\mu}) = -\frac{1}{2}\sigma_{XY}\sigma_X^{-1}\sigma_Y^{-3}$$

and so from Lemma 5.9

$$\tau_{XYr}^{\sim}(M_i) = \text{corr}\left[h_r(X_{ij},Y_{ij}),h_r(X_{ij'},Y_{ij'})\middle|M_i\right]$$

where

$$h_r(X_{ij},Y_{ij}) = (X_{ij}-\mu_X)(Y_{ij}-\mu_Y)\sigma_X^{-1}\sigma_Y^{-1}$$

$$+ (X_{ij}-\mu_X)^2(-\tfrac{1}{2}\sigma_{XY}\sigma_X^{-3}\sigma_Y^{-1})$$

$$+ (Y_{ij}-\mu_Y)^2(-\tfrac{1}{2}\sigma_{XY}\sigma_X^{-1}\sigma_Y^{-3})$$

as required.

Lemma 7.2

If B holds

$$\tau_{XYr}^{\sim}(M_i) \doteq V_I(r_i|M_i)/m_o V_{II}(T_{XYr}^{\sim}|s,\underline{M})$$

where

$$r_i = \frac{(\mu_{Xi}-\mu_X)(\mu_{Yi}-\mu_Y)+\sigma_{XYi}}{\left[(\mu_{Xi}-\mu_X)^2+\sigma_{Xi}^2\right]^{\frac{1}{2}}\left[(\mu_{Yi}-\mu_Y)^2+\sigma_{Yi}^2\right]^{\frac{1}{2}}} \qquad (7.8)$$

Proof:

From Lemma 5.10 and (7.7) we have

$$\tau_{XYr}^{\sim}(M_i) \doteq V_I\left[g(U_{i1},U_{i2},U_{i3})|M_i\right]/m_o V_{II}(T_{XYr}^{\sim}|s,\underline{M})$$

where

$$U_{i1} = m_o E_I\left(h_c(X_{ij},Y_{ij})|\theta_i\right)$$

$$= m_o\left[(\mu_{Xi}-\mu_X)(\mu_{Yi}-\mu_Y) + \sigma_{XYi}\right]/(m_o-1)$$

$$U_{i2} = m_o\, E_I\left(h_v(X_{ij})|\theta_i\right)$$

$$= m_o\left[(\mu_{Xi}-\mu_X)^2 + \sigma_{Xi}^2\right]/(m_o-1)$$

$$U_{i3} = m_o\, E_I\left(h_v(Y_{ij})|\theta_i\right)$$

$$= m_o\left[(\mu_{Yi}-\mu_Y)^2 + \sigma_{Yi}^2\right]/(m_o-1)$$

Hence

$$g(U_{i1},U_{i2},U_{i3}) = r_i$$

as required.

The quantity r_i is of fundamental importance in determining τ_{XY_r} . r_i may be viewed as a generalised within-cluster correlation between X and Y such that $E(r_i) \doteq \rho$. The following Lemma shows that r_i is not greater than unity in absolute value.

Lemma 7.3

For r_i defined in (7.8)

$$-1 \leq r_i \leq 1$$

Proof Let $\alpha = \mu_{Xi} - \mu_X$ and $\beta = \mu_{Yi} - \mu_Y$

Then

$$1 - r_i^2 = K\left[(\alpha^2+\sigma_{Xi}^2)(\beta^2+\sigma_{Yi}^2)-(\alpha\beta+\sigma_{XYi})^2\right]$$

where

$$K = (\alpha^2 + \sigma_{Xi}^2)^{-1}(\beta^2 + \sigma_{Yi}^2)^{-1} \geq 0$$

Hence $1 - r_i^2 = K\left[\alpha^2\sigma_{Yi}^2 + \beta^2\sigma_{Xi}^2-2\alpha\beta\sigma_{XYi} + (\sigma_{Xi}^2\sigma_{Yi}^2-\sigma_{XYi}^2)\right]$

$$\geq K\left[\alpha^2\sigma_{Yi}^2 + \beta^2\sigma_{Xi}^2 - 2|\alpha\beta||\sigma_{XYi}|\right]$$

$$\geq K\left[\alpha^2\sigma_{Yi}^2 + \beta^2\sigma_{Xi}^2 - 2|\alpha\beta|\sigma_{Xi}\sigma_{Yi}\right]$$

$$= K(|\alpha|\sigma_{Yi} - |\beta|\sigma_{Xi})^2$$

$$\geq 0$$

If $\mu_{Xi} = \mu_X$ and $\mu_{Yi} = \mu_Y$ then $r_i = \rho_i = \sigma_{XYi}/\sigma_{Xi}\sigma_{Yi}$ and the intra-cluster correlation τ_{XY_r} is a measure of the variation in the cluster correlations, ρ_i. In general r_i is not equal to ρ_i. In Figure 7.1 contours of constant r_i are plotted for values of $\mu_{Xi} - \mu_X$ and $\mu_{Yi} - \mu_Y$ where $\rho_i = 0$ and $\sigma_{Xi}^2 = \sigma_{Yi}^2 = 1$. In Figure 7.2 the same contours are plotted for $\rho_i = 0.5$. We may make two observations from these Figures.

Firstly, if τ_{Xm} and τ_{Ym} are low then r_i is approximately equal to ρ_i. For example, suppose $\tau_{Xm} \doteq 0.02$, $\tau_{Ym} \doteq 0.02$ as in the Family Expenditure Survey data. Then, assuming $\sigma_W^2 = 1$ the

Figure 7.1 : r_i contours for $\rho_i = 0$ ($\sigma_{Xi}^2 = \sigma_{Yi}^2 = 1$)

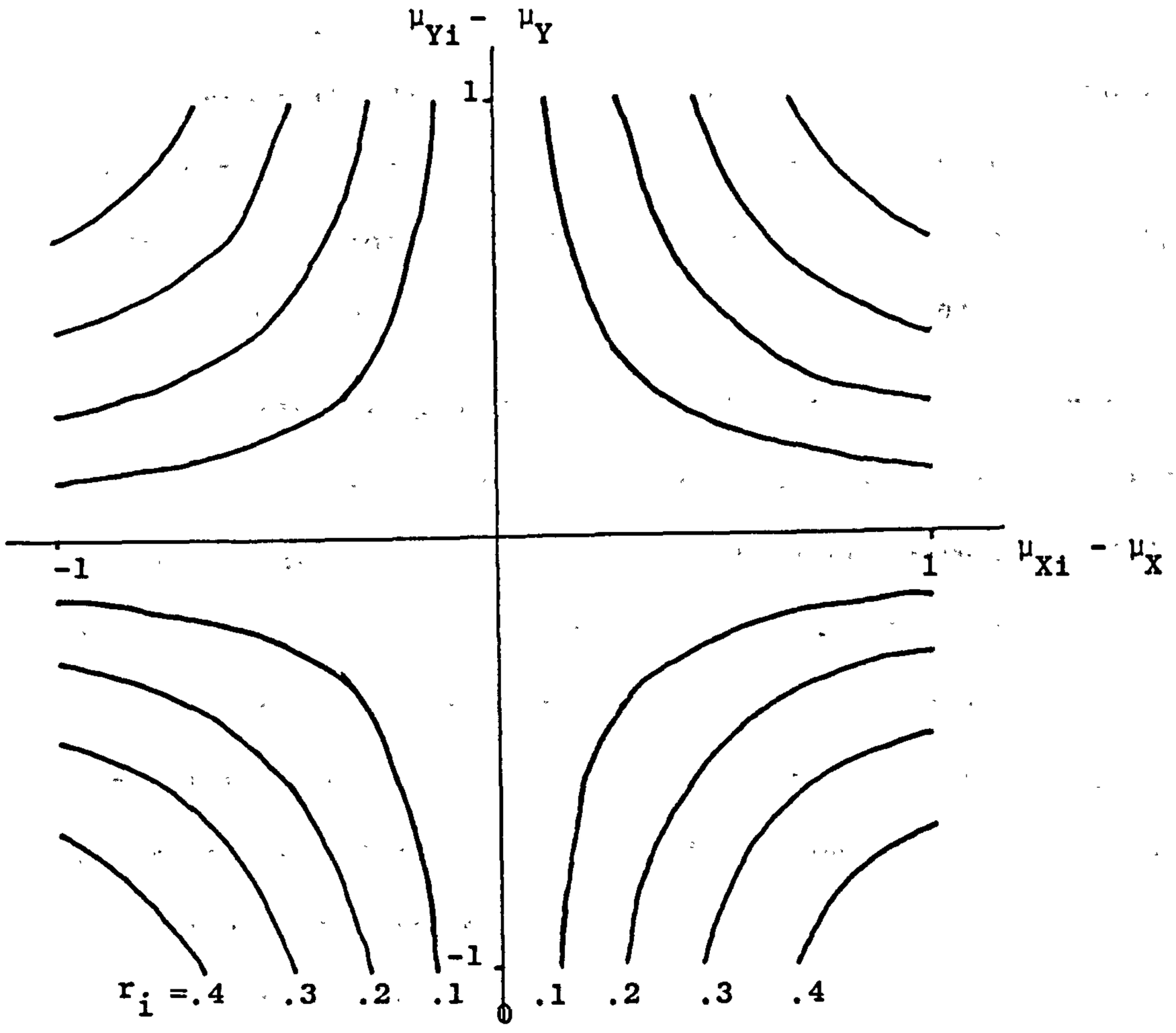
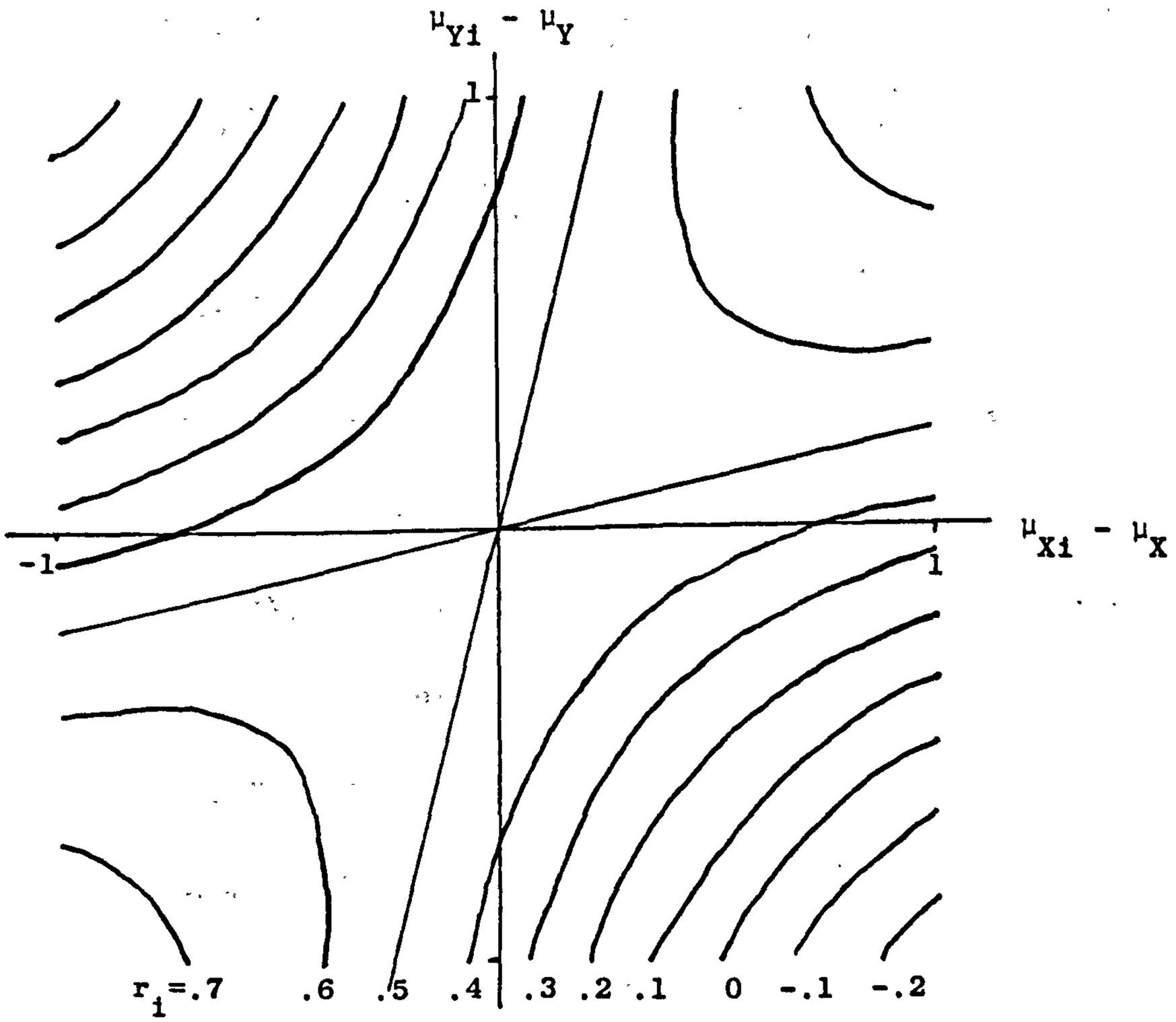


Figure 7.2 : r_i contours for $\rho_i = 0.5$ ($\sigma_{Xi}^2 = \sigma_{Yi}^2 = 1$)



standard deviation of $\mu_{Xi} - \mu_X$ (or $\mu_{Yi} - \mu_Y$) is approximately 0.14. Hence, by inspecting the Figures, in at least 95% of the clusters we might expect the difference between r_i and ρ_i to be at most 0.1, and for most clusters rather less than this. In contrast, the standard deviation of the between-cluster distribution of ρ_i is approximately 0.15 for the three FES variables.

Secondly, we note from Figure 7.2 that, for given values of τ_{Xm} and τ_{Ym}, the expected difference between r_i and ρ_i is smaller if the within-correlation ρ_i and the between correlation (of μ_{Xi} and μ_{Yi}) are of the same sign and larger if they are of opposite sign.

The following two Lemmas permit us to express $\tau_{XY\tilde{r}}$ in terms of population moments and hence to examine the form of $\tau_{XY\tilde{r}}$ in some special cases. First of all we define some more notation, under the assumption that B holds.

Let $k_{31} = E_{II}(X_{ij} - \mu_X)^3(Y_{ij} - \mu_Y) - 3\,\sigma_X^2\,\sigma_{XY}$

$k_{13} = E_{II}(X_{ij} - \mu_X)(Y_{ij} - \mu_Y)^3 - 3\,\sigma_Y^2\,\sigma_{XY}$

$k_{31B}(M_i) = E_I[(\mu_{Xi} - \mu_X)^3(\mu_{Yi} - \mu_Y)|M_i] - 3\,\sigma_{XB}^2(M_i)\sigma_{XYB}(M_i)$

$k_{13B}(M_i) = E_I[(\mu_{Xi} - \mu_X)(\mu_{Yi} - \mu_Y)^3|M_i] - 3\,\sigma_{YB}^2(M_i)\sigma_{XYB}(M_i)$

$c_{X\cdot Y}(M_i) = \mathrm{cov}_I[\sigma_{Xi}^2,\ (\mu_{Yi} - \mu_Y)^2|M_i]$

$c_{Y\cdot X}(M_i) = \mathrm{cov}_I[\sigma_{Yi}^2,\ (\mu_{Xi} - \mu_Y)^2|M_i]$

$c_{X\cdot XY}(M_i) = \mathrm{cov}_I[\sigma_{Xi}^2,\ (\mu_{Xi} - \mu_X)(\mu_{Yi} - \mu_Y)|M_i]$

$c_{XY\cdot X}(M_i) = \mathrm{cov}_I[\sigma_{XYi},\ (\mu_{Xi} - \mu_X)^2|M_i]$

$c_{Y\cdot XY}(M_i) = \mathrm{cov}_I[\sigma_{Yi}^2,\ (\mu_{Xi} - \mu_X)(\mu_{Yi} - \mu_Y)|M_i]$

$c_{XY\cdot Y}(M_i) = \mathrm{cov}_I[\sigma_{XYi},\ (\mu_{Yi} - \mu_Y)^2|M_i]$

$\delta_{X\cdot Y}(M_i) = \mathrm{cov}_I[\sigma_{Xi}^2,\ \sigma_{Yi}^2|M_i]$

$\delta_{X\cdot XY}(M_i) = \mathrm{cov}_I[\sigma_{Xi}^2,\ \sigma_{XYi}|M_i]$

$\delta_{Y\cdot XY}(M_i) = \mathrm{cov}_I[\sigma_{Yi}^2,\ \sigma_{XYi}|M_i]$

Lemma 7.4

$$V_{II}(T_{XYr}^{\sim}|s,\underline{M}) \doteq \left[(1-\rho^2)^2 + \rho^2 K\right]/m_o \qquad (7.9)$$

where $K = k_{22}(\sigma_{XY}^{-2} + \sigma_X^{-2}\sigma_Y^{-2}/2) + k_{4X}/4\ \sigma_X^{4} + k_{4Y}/4\ \sigma_Y^{4}$

$$- k_{31}/\sigma_X^{2}\sigma_{XY} - k_{13}/\sigma_Y^{2}\sigma_{XY}$$

and k_{4X} and k_{4Y} are defined in Lemma 5.16

If the joint distribution of (X_{ij}, Y_{ij}) is bivariate normal then $K = 0$.

Proof: The first order Taylor Series expansion of $\text{var}_{II}[T_{XYr}^{\sim}]$ is

(c.f. Kendall and Stuart, 1969, Ch. 10)

$$\text{var}_{II}(T_{XYr}^{\sim}) = (m_o - 1)^2 \rho^2 \left[\text{var}_{II}(T_{XYc}^{\sim})/\sigma_{XY}^{2} + \left[\text{var}_{II}(T_{X\tilde{V}})/\sigma_X^{4}\right.\right.$$

$$+ \text{var}_{II}(T_{Y\tilde{V}})/\sigma_Y^{4} + 2\ \text{cov}_{II}(T_{X\tilde{V}}, T_{Y\tilde{V}})/\sigma_X^{2}\sigma_Y^{2}\left.\right]/4$$

$$\left. - \text{cov}_{II}(T_{XYc}^{\sim}, T_{X\tilde{V}})/\sigma_{XY}\sigma_X^{2} - \text{cov}_{II}(T_{XYc}^{\sim}, T_{Y\tilde{V}})/\sigma_{XY}\sigma_Y^{2}\right]/m_o^{2} \qquad (7.10)$$

From the proofs of Lemmas 5.17 and 5.27

$$\text{var}_{II}(T_{XYc}^{\sim}) = m_o(\sigma_X^{2}\sigma_Y^{2} + \sigma_{XY}^{2} + k_{22})/(m_o - 1)^2 \qquad (7.11)$$

$$\text{var}_{II}(T_{X\tilde{V}}) = m_o(2\ \sigma_X^{4} + k_{4X})/(m_o - 1)^2 \qquad (7.12)$$

$$\text{var}_{II}(T_{Y\tilde{V}}) = m_o(2\ \sigma_Y^{4} + k_{4Y})/(m_o - 1)^2 \qquad (7.13)$$

Now

$$\text{cov}_{II}(T_{X\tilde{V}}, T_{Y\tilde{V}}) = m_o\ \text{cov}_{II}\left[(X_{ij} - \mu_X)^2, (Y_{ij} - \mu_Y)^2\right]/(m_o - 1)^2$$

$$= m_o(2\ \sigma_{XY}^{2} + k_{22})/(m_o - 1)^2 \qquad (7.14)$$

$$\text{cov}_{II}(T_{XYc}^{\sim}, T_{X\tilde{V}}) = m_o\ \text{cov}_{II}\left[(X_{ij} - \mu_X)(Y_{ij} - \mu_Y), (X_{ij} - \mu_X)^2\right]/(m_o - 1)^2$$

$$= m_o(2\ \sigma_{XY}\ \sigma_X^{2} + k_{31})/(m_o - 1)^2 \qquad (7.15)$$

Similarly

$$\text{cov}_{II}(T_{XY}^{\sim}, T_{YV}^{\sim}) = m_o(2\ \sigma_{XY}\ \sigma_Y^{2} + k_{13})/(m_o - 1)^2 \qquad (7.16)$$

Substituting (7.11) - (7.16) into (7.10) gives

$$\text{var}_{II}(T_{XYr}^{\sim}) \doteq \rho^2\left[(1+\rho^2)/\rho^2 + (2 + 2 + 4\rho^2)/4 - 2 - 2\right.$$

$$\left. + k_{22}/\sigma_{XY}^{2} + (k_{4X}/\sigma_X^{4} + k_{4Y}/\sigma_Y^{4} + 2\ k_{22}/\sigma_X^{2}\sigma_Y^{2})/4\right.$$

$$- k_{31}/\sigma_{XY}\sigma_X^2 - k_{13}/\sigma_{XY}\sigma_Y^2]/m_o$$

$$= ((1-\rho^2)^2 + \rho^2 K)/m_o \text{ as required.}$$

Lemma 7.5

If B holds

$$\text{var}_I(r_i|M_i) \doteq \rho^2[A_1/\sigma_{XY}^2 + (A_2/\sigma_X^4 + A_3/\sigma_Y^4 + 2A_4/\sigma_X^2\sigma_Y^2)/4$$

$$- A_5/\sigma_{XY}\sigma_X^2 - A_6/\sigma_{XY}\sigma_Y^2] \qquad (7.17)$$

where $A_1 = \text{var}_I[(\mu_{Xi}-\mu_X)(\mu_{Yi}-\mu_Y) + \sigma_{XYi}|M_i]$

$$= \sigma_{XB}^2(M_i)\,\sigma_{YB}^2(M_i) + \sigma_{XYB}^2(M_i) + k_{22B}(M_i) + 2c_{1XY}(M_i)$$

$$+ \gamma_{XY}(M_i)$$

$A_2 = \text{var}_I[(\mu_{Xi}-\mu_X)^2 + \sigma_{Xi}^2|M_i]$

$= 2\sigma_{XB}^4(M_i) + k_{4XB}(M_i) + 2c_{1X}(M_i) + \gamma_X(M_i)$

$A_3 = \text{var}_I[(\mu_{Yi}-\mu_Y)^2 + \sigma_{Yi}^2|M_i]$

$= 2\sigma_{YB}^4(M_i) + k_{4YB}(M_i) + 2c_{1Y}(M_i) + \gamma_Y(M_i)$

$A_4 = \text{cov}_I[(\mu_{Xi}-\mu_X)^2 + \sigma_{Xi}^2\ ,\ (\mu_{Yi}-\mu_Y)^2 + \sigma_{Yi}^2|M_i]$

$= 2\sigma_{XYB}^2(M_i) + k_{22B}(M_i) + c_{X\cdot Y}(M_i) + c_{Y\cdot X}(M_i) + \delta_{X\cdot Y}(M_i)$

$A_5 = \text{cov}_I[(\mu_{Xi}-\mu_X)(\mu_{Yi}-\mu_Y) + \sigma_{XYi}\ ,\ (\mu_{Xi}-\mu_X)^2 + \sigma_{Xi}^2|M_i]$

$= 2\sigma_{XYB}(M_i)\sigma_{XB}^2(M_i) + k_{31B}(M_i) + c_{XY\cdot X}(M_i) + c_{X\cdot XY}(M_i)$

$$+ \delta_{X\cdot XY}(M_i)$$

$A_6 = \text{cov}_I[(\mu_{Xi}-\mu_X)(\mu_{Yi}-\mu_Y) + \sigma_{XYi}\ ,\ (\mu_{Yi}-\mu_Y)^2 + \sigma_{Yi}^2|M_i]$

$= 2\sigma_{XYB}(M_i)\sigma_{YB}^2(M_i) + k_{13B}(M_i) + c_{XY\cdot Y}(M_i) + c_{Y\cdot XY}(M_i)$

$$+ \delta_{X\cdot XY}(M_i)$$

Proof : (7.17) follows as in Lemma 7.4 from the first-order Taylor Series expansion of $var_I(r_i|M_i)$ and by using the results of Lemmas 5.17 and 5.27.

We now use the results of Lemmas 7.2, 7.4 and 7.5 to evaluate $\tau_{XYr}^{\sim}$ for some special cases.

Case 1 : A holds, $\sigma_{Xi}^2 = \sigma_{XW}^2$, $\sigma_{Yi}^2 = \sigma_{YW}^2$, $\sigma_{XYi} = \sigma_{XYW}$ (common within-cluster covariance matrix)

In this case

$$c_{X\cdot Y}(M_i) = c_{Y\cdot X}(M_i) = c_{X\cdot XY}(M_i) = c_{XY\cdot X}(M_i)$$
$$= c_{Y\cdot XY}(M_i) = c_{XY\cdot Y}(M_i) = \sigma_{X\cdot Y}(M_i) = \delta_{X\cdot XY}(M_i)$$
$$= \delta_{Y\cdot XY}(M_i) = \gamma_X(M_i) = \gamma_Y(M_i) = \gamma_{XY}(M_i)$$
$$= c_{1X}(M_i) = c_{1Y}(M_i) = c_{1XY}(M_i) = 0$$

And since A holds, susbtituting into Lemma 7.5 we have

$$A_1 = \sigma_{XB}^2 \sigma_{YB}^2 + \sigma_{XYB}^2 + k_{22B}$$

$$A_2 = 2\sigma_{XB}^4 + k_{4XB} \qquad A_3 = 2\sigma_{YB}^4 + k_{4YB}$$

$$A_4 = 2\sigma_{XYB}^2 + k_{22B}$$

$$A_5 = 2\sigma_{XYB}\sigma_{XB}^2 + k_{31B} \qquad A_6 = 2\sigma_{XYB}\sigma_{YB}^2 + k_{13B}$$

Hence, substituting into Lemmas 7.2, 7.4 and 7.5

$$\tau_{XYr}^{\sim} = \left[\tau_{Xm}\tau_{Ym}(1 + \rho_B^2) + \tau_{Xm}^2\rho^2/2 + \tau_{Ym}^2\rho^2/2 + \tau_{Xm}\tau_{Ym}\rho_B^2\rho^2 - 2\tau_{Xm}\sqrt{\tau_{Xm}\tau_{Ym}}\,\rho_B\,\rho - 2\tau_{Ym}\sqrt{\tau_{Xm}\tau_{Ym}}\,\rho_B\,\rho + \rho^2K'\right] / \left[(1-\rho^2)^2 + \rho^2K\right] \qquad (7.18)$$

where $K' = k_{22B}/\sigma_{XY}^2 + (k_{4XB}/\sigma_X^4 + k_{4YB}/\sigma_Y^4 + 2k_{22B}/\sigma_X^2\sigma_Y^2)/4 - 2k_{31B}/\sigma_{XY}\sigma_X^2 - 2k_{13B}/\sigma_{XY}\sigma_Y^2$

We note that $\tau_{XY\tilde{r}}$ is a 'quadratic' function of τ_{Xm} and τ_{Ym}. Hence if (i) both τ_{Xm} and τ_{Ym} are small and (ii) the between cluster distribution of μ_{Xi} and μ_{Yi} is close to joint normality (to that K' is very small) then $\tau_{XY\tilde{r}}$ is very small.

Note, however, that it does not follow, as for $\tau_{XY\tilde{c}}$, that if either τ_{Xm} or τ_{Ym} is small then $\tau_{XY\tilde{r}}$ is small. For if $\tau_{Ym} = 0$. Then

$$\tau_{XY\tilde{r}} = \rho^2(\tau_{Xm}^2/2 + K')/[(1 - \rho^2)^2 + \rho^2 K]$$

Case 2: A holds, $\mu_{Yi} = \mu_Y$, $\sigma_{Yi}^2 = \sigma_{Yw}^2$ (no clustering on Y)

In this case

$$A_1 = \gamma_{XY} \qquad A_2 = 2\sigma_{XB}^4 + k_{4XB} + 2c_{1X} + \gamma_X$$

$$A_3 = A_4 = A_6 = 0 \qquad A_5 = c_{XY\cdot X} + \delta_{X\cdot XY}$$

Hence, substituting into Lemmas 7.2, 7.4 and 7.5

$$\tau_{XY\tilde{r}} = \rho^2[\gamma_{XY}/\sigma_{XY}^2 + (2\sigma_{XB}^2 + k_{4XB} + 2c_{1X} + \gamma_X)/4\sigma_X^4 - (c_{XY\cdot X} + \delta_{X\cdot XY})/\sigma_{XY}\sigma_X^2]/(1 - \rho^2)^2 + \rho^2 K)$$

So, as for $\tau_{XY\tilde{c}}$, $\tau_{XY\tilde{r}}$ is not necessarily zero in this case. However, unlike $\tau_{XY\tilde{c}}$, even if $\sigma_{XYi} = \sigma_{XYW}$ is constant then $\tau_{XY\tilde{r}}$ is in general non-zero since in this case

$$\tau_{XY\tilde{r}} = \frac{\rho^2(2\sigma_X^4 + k_{4X})\ \tau_{X\tilde{v}}/4\sigma_X^4}{(1 - \rho^2)^2 + \rho^2 K} \qquad \text{from (5.45)}$$

$$= \frac{\rho^2\tau_{X\tilde{v}}}{(1 - \rho^2)^2}$$

if the marginal distribution of X is normal.

In contrast to these formulae, Bebbington and Smith (1977) suggest that

$$\tau_{XYr} \propto \min(\tau_{Xm}, \tau_{Ym})$$

provides 'quite a good predictive equation', on the basis of their

empirical evidence. For the present case this formula suggests that τ_{XYr} should be zero.

Case 3: A holds, $\mu_{Xi} = \mu_X$, $\mu_{Yi} = \mu_Y$ (common within-cluster means)

In this case

$$r_i = \rho_i = \sigma_{XYi}/\sigma_{Xi}\sigma_{Yi}$$

and so from (6.26) and (6.28)

$$\tau_{XY\tilde{r}} = \mathrm{var}_I(\rho_i)/[(1-\rho^2)^2 + \rho^2 K] \qquad (7.19)$$

whereas τ_{Xm} and τ_{Ym} are zero. This is, of course, in contrast to Kish and Frankel's (1974) conjectures.

Case 4: A holds

In general, if A holds, $\tau_{XY\tilde{r}}$ will be a combination of expressions (7.18) and (7.19) We may write

$$\tau_{XY\tilde{r}} = \tau_1 + \tau_2 + \tau_3$$

where τ_1 is given by the expression for $\tau_{XY\tilde{r}}$ in (7.18), τ_3 is given by (7.19) and τ_2 is such that

$$|\tau_2| \leq 2(\tau_1\tau_3)^{\frac{1}{2}}$$

Some estimates of these quantities, under the assumption that X_{ij} and Y_{ij} are normally distributed within clusters are given in Table 7.1

Table 7.1 Estimates for FES Data

Variables X, Y	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_{XY\tilde{r}}$
log(V1), log(V2)	.0003	-.003	.061	.0577
log(V1), log(V3)	.0005	-.011	.079	.0683
log(V2), log(V3)	.0004	-.003	.070	.0673

We note firstly that $\hat{\tau}_{XY\tilde{r}}$ is dominated by $\hat{\tau}_3$ as in Table 5.4 for $\tau_{XY\tilde{c}}$. So the variation in the cluster correlation ρ_i seems

to account for most of the intra cluster correlation. We might expect this to be typically the case when both τ_{Xm} and τ_{Ym} are small (and so τ_1 is very small). We note secondly that the values of $\hat{\tau}_{XYr}^{\sim}$ in Table 7.1 are consistently much larger than the corresponding values of $\hat{\tau}_{XYc}^{\sim}$ in Table 5.4. Indeed, although all the values of $\hat{\tau}_{XYc}^{\sim}$ are less than all the values of τ_{Ym} in Table 5.2 in accordance with Kish and Frankel's (1974) conjectures, the values of $\hat{\tau}_{XYr}^{\sim}$ are uniformly <u>greater</u> than all the values of $\hat{\tau}_{Ym}$. This differs from the empirical results of Frankel (1971). One explanation for this fact is that the values of ρ are larger here than is 'common' in social surveys ($\hat{\rho}_{12} = .68$, $\hat{\rho}_{13} = .81$, $\hat{\rho}_{23} = .77$) and therefore $(1 - \rho^2)^2$, the dominating term in the denominator of $\tau_{XYr}^{\sim}$, is unusually small.

<u>Case 5 : B holds but A does not hold</u>

As before we just consider the spatial process approach of Sections 5.3 - 5.5. As for $\tau_{XYc}^{\sim}$ we assume (X,Y) follows a stationary Gaussian isotropic spatial process. In the notation of Section 5.3 r_i, as defined in (7.8) , may be written

$$r_i = \frac{\int_{\Omega i} h_3(X(\underline{x}), Y(\underline{x}))d\underline{x}}{\left[\int_{\Omega i} h_2(X(\underline{x}))d\underline{x} \int_{\Omega i} h_2(Y(\underline{x}))d\underline{x}\right]^{\frac{1}{2}}}$$

Using the fact that

$$\mathrm{var}_I\left[(m_o-1)h_v(X(\underline{x}))\right] = 2\sigma_X^4$$

$$\mathrm{var}_I\left[(m_o-1)h_v(Y(\underline{x}))\right] = 2\sigma_Y^4$$

and $$\mathrm{var}_I\left[(m_o-1)h_c(X(\underline{x}), Y(\underline{x}))\right] = (1+\rho^2)\sigma_X^2\sigma_Y^2$$

we may obtain the first-order Taylor series expansion of $\mathrm{var}_I(r_i|M_i)$ as

$$\begin{aligned}\mathrm{var}_I(r_i|M_i) = \rho^2\Big[&(1+\rho^2)\tau_{XYci}^{\sim}/\rho^2 + \tfrac{1}{4}\,(2\tau_{Xv}^{\sim} + 2\tau_{Yv}^{\sim} \\ &+ 4\iint \mathrm{corr}_I\big[h_2(X(\underline{x})),\ h_2(Y(\underline{x}'))\big]d\underline{x}\ d\underline{x}' \\ &- \frac{\sqrt{2(1+\rho^2)}}{\rho}\iint \mathrm{corr}_I\big[h_2(X(\underline{x})),\ h_3(X(\underline{x}'),Y(\underline{x}'))\big]d\underline{x}d\underline{x}' \\ &- \frac{\sqrt{2(1+\rho^2)}}{\rho}\iint \mathrm{corr}_I\big[h_2(Y(\underline{x})),\ h_3(X(\underline{x}'),Y(\underline{x}'))\big]d\underline{x}d\underline{x}'\Big]\end{aligned} \tag{7.20}$$

Using standard results for the fourth moments of multivariate normal distributions

$$\mathrm{corr}_I\big[h_{\dot{v}}(X(\underline{x})),\ h_{\dot{v}}(Y(\underline{x}'))\big] = r_{XY}(s)^2$$

$$\mathrm{corr}_I\big[h_v(X(\underline{x})),\ h_c(X(\underline{x}'),\ Y(\underline{x}'))\big] = \frac{2}{\sqrt{(1+\rho^2)}}\, r_X(s)r_{XY}(s)$$

$$\mathrm{corr}_I\big[h_v(Y(\underline{x}),\ h_c(X(\underline{x}'),\ Y(\underline{x}')\big] = \frac{2}{\sqrt{2(1+\rho^2)}}\, r_Y(s)r_{XY}(s)$$

Substituting these expression into (7.20) and using (5.56) - (5.58) and (5.98) we obtain

$$\begin{aligned}\mathrm{var}_I(r_i|M_i) = \int_o^1\Big[&2(1+\rho^2)\ r_{XY}{}^2(\alpha_i t) + 2r_X(\alpha_i t)r_Y(\alpha_i t) \\ &+ \rho^2 r_X{}^2(\alpha_i t) + \rho^2 r_Y{}^2(\alpha_i t) - 4\rho r_{XY}(\alpha_i t)(r_X(\alpha_i t) \\ &+ r_Y(\alpha_i t))\Big]\ K(\alpha_i t)\ dt/2 \qquad \text{where } \alpha_i = kM_i^{\frac{1}{2}}\end{aligned}$$

Hence from Lemmas 7.2 and 7.4

$$\begin{aligned}\tau_{XYr}^{\sim}(M_i) = \int_o^1\Big[&2(1+\rho^2)\ r_{XY}{}^2(\alpha_i t) + 2r_X(\alpha_i t)\ r_Y(\alpha_i t) \\ &+ \rho^2 r_X{}^2(\alpha_i t) + \rho^2\ r_Y{}^2(\alpha_i t) - 4\rho r_{XY}\ (\alpha_i t)(r_X(\alpha_i t) \\ &+ r_Y(\alpha_i t))\Big]\ K(\alpha_i t)\ dt/2(1-\rho^2)^2\end{aligned}$$

This may be compared with (7.18). As in (5.98) $\tau_{XYr}^{\sim}$ is an integral of a quadratic function of $r_X(\alpha_i t)$ and $r_{XY}(\alpha_i t)$. However, $\tau_{XYc}^{\sim}$ is small if either r_X or r_Y is small whereas $\tau_{XYr}^{\sim}$ is only small if both r_X and r_Y are small. In the same way the rate of decay of $\tau_{XYc}^{\sim}$ as a function of M_i is 'determined' by the rate of decay of the faster of τ_{Xm} and τ_{Ym} whereas the rate of decay of $\tau_{XYr}^{\sim}$ is determined by the slower of the two. Expression for $\tau_{XYr}^{\sim}$ as a function of M_i may be obtained by substituting particular functional forms for $K(\cdot)$.

7.2 Regression Coefficients

As can be seen from the table in Section 1.5 there is a sizeable literature on the regression analysis of clustered survey data. This literature divides broadly into (i) variance estimation of the standard OLS estimator (Frankel, 1971; Kish and Frankel, 1974; Fuller, 1975; Shah et al, 1977) and (ii) estimation of the parameters of a regression model reflecting the clustered population structure (Konijn, 1962; Porter, 1973; Campbell, 1977; Pfefferman and Nathan, 1981; Holt and Scott, 1981). In the terminology of Section 1.3.2, we would refer to the latter work as being concerned with disaggregated targets of inference and, as such, not of relevance to us. In specific cases, however, our models will coincide. Similarly, the work on variance estimation is of little relevance since it does not attempt to investigate the theoretical properties of the OLS estimator. One exception is Fuller (1975) who suggests an extension of a result similar to our Lemma 5.11b to that of the OLS estimator.

In this section we consider the estimation of $\beta = \sigma_{XY}/\sigma_X^2$ by the standard estimator

$$T_{XYb} = T_{XYc}/T_{Xv} \tag{7.21}$$

As in Section 7.1 we suppose that Assumption B holds and that T_{XYb} is approximately unbiased for β. In Frankel (1971) the average bias of T_{XYb} with $n = 12$ was about 5% of the average standard error. We shall approximate the misspecification effect of T_{XYb} by that of

$$T_{XY\tilde{b}} = T_{XY\tilde{c}}/T_{X\tilde{v}} \tag{7.22}$$

Lemma 7.6

If Assumption B holds

$$\mathrm{meff}(T_{XYb}|s,\underline{M}) \doteq 1 + \sum_{i=1}^{n} m_i(m_i-1)\tau_{XYb}\tilde{}(M_i)/m_o \tag{7.23}$$

where

$$\tau_{XY\tilde{b}}(M_i) = \text{corr}_I\left[h_b(X_{ij},Y_{ij}),\ h_b(X_{ij'},Y_{ij'})\,\middle|\,M_i\right] \quad j \neq j' \tag{7.24}$$

$$h_b(X,Y) = (X-\mu_X)\left(Y-\mu_Y-\beta(X-\mu_X)\right) \tag{7.25}$$

Proof:

This follows from Lemmas 5.5 and 5.9 noting that

$$T_{XY\tilde{b}} = g(T_{XY\tilde{c}},\ T_{X\tilde{v}})$$

where

$$g(x,y) = x/y$$

$$g_x(m_o\underline{\mu}) = 1/\sigma_x^2$$

$$g_y(m_o\underline{\mu}) = -\sigma_{XY}/\sigma_X^4$$

∴ from Lemma 5.9

$$\tau_{XY\tilde{b}}(M_i) = \text{corr}_I\left[h_b(X_{ij},Y_{ij}),\ h_b(X_{ij'},Y_{ij'})\,\middle|\,M_i\right]$$

where

$$h_b(X,Y) = (X-\mu_X)(Y-\mu_Y)/\sigma_X^2-(X-\mu_X)^2\sigma_{XY}/\sigma_X^4 \propto (X-\mu_X)\left(Y-\mu_Y-\beta(X-\mu_X)\right)$$

Note that $h_b(X_{ij},Y_{ij}) = (X_{ij}-\mu_X)e_{ij}$ where e_{ij} is the regression 'residual' $e_{ij} = Y_{ij}-\mu_Y-\beta(X_{ij}-\mu_X)$. This suggests a variance estimation procedure based on $\hat{h}_b(X_{ij},Y_{ij}) = (X_{ij}-\hat{\mu}_X)\left(Y_{ij}-\hat{\mu}_Y-\hat{\beta}(X_{ij}-\hat{\mu}_X)\right)$ as in Fuller (1975, p.123).

Lemma 7.7

If B holds

$$\tau_{XY\tilde{b}}(M_i) \doteq V_I(b_i|M_i)/m_oV_{II}(T_{XY\tilde{b}}|s,\underline{M}) \tag{7.26}$$

where

$$b_i = \frac{(\mu_{Xi}-\mu_X)(\mu_{Yi}-\mu_Y) + \sigma_{XYi}}{(\mu_{Xi}-\mu_X)^2 + \sigma_{Xi}^2} \tag{7.27}$$

Proof:

By analogy with the proof of Lemma 7.2.

Lemma 7.8

$$V_{II}(T_{XY\tilde{b}}|s,\underline{M}) \doteq \sigma_Y^2(1-\rho^2 + K)/\sigma_X^2 m_o \qquad (7.28)$$

where $K = k_{22}/\sigma_X^2\sigma_Y^2 + k_{4X}\,\sigma_{XY}^2/\sigma_X^6\,\sigma_Y^2 - 2k_{31}\sigma_{XY}/\sigma_X^4\sigma_Y^2$

(Note: if the joint distribution of (X_{ij},Y_{ij}) under Model II is normal then K=0)

Proof : The first-order Taylor series expansion of $var_{II}(T_{XY\tilde{b}})$ is

$$var_{II}(T_{XY\tilde{b}}) \doteq (m_o-1)^2\beta^2\left[var_{II}(T_{XY\tilde{c}})/\sigma_{XY}^2 + var_{II}(T_{X\tilde{v}})/\sigma_X^4 - 2cov_{II}(T_{XY\tilde{c}}, T_{XY\tilde{v}})/\sigma_{XY}\sigma_X^2\right]/m_o^2$$

Substituting from (7.11), (7.12) and (7.15)

$$var_{II}(T_{XY\tilde{b}}) \doteq \beta^2\left[(\sigma_X^2\sigma_Y^2 + \sigma_{XY}^2 + k_{22})/\sigma_{XY}^2 + (2\sigma_X^4+k_{4X})/\sigma_X^4 - 2(2\sigma_{XY}\sigma_X^2 + k_{31})/\sigma_X^2\sigma_{XY}\right]/m_o$$

which gives (7.28).

Lemma 7.9

If B holds

$$var_I(b_i|M_i) \doteq \beta^2\left[A_1/\sigma_{XY}^2 + A_2/\sigma_X^4 - 2A_5/\sigma_{XY}\sigma_X^2\right] \qquad (7.29)$$

where A_1, A_2 and A_5 are given in Lemma 7.5.

Proof

(7.29) is the first-order Taylor series expansion of $var_I(b_i|M_i)$

We now use the results of Lemmas 7.7 - 7.9 to evaluate $\tau_{XY\tilde{b}}$ for some special cases. If Assumption A holds we define the between-cluster and mean within-cluster regression coefficients as

$$\beta_B = \sigma_{XYB}/\sigma_{XB}^2$$

$$\beta_W = \sigma_{XYW}/\sigma_{XW}^2$$

respectively. We note that β is a weighted mean of β_B and β_W.

$$\beta = \tau_{Xm}\beta_B + (1-\tau_{Xm})\beta_W \qquad (7.30)$$

Since τ_{Xm} is usually small β will usually be very close to β_W.

Case 1 : A holds, $\sigma_{Xi}^2 = \sigma_{XW}^2$, $\sigma_{Yi}^2 = \sigma_{YW}^2$, $\sigma_{XYi} = \sigma_{XYW}$, $\beta_B = \beta_W = \beta$

In this case

$$A_1 = \sigma_{XB}^2\,\sigma_{YB}^2 + \sigma_{XYB}^2 + k_{22B}$$

$$A_2 = 2\sigma_{XB}^4 + k_{4XB}$$

$$A_5 = 2\sigma_{XYB}\,\sigma_{XB}^2 + k_{31B}$$

and so from Lemma 7.9

$$\mathrm{var}_I(b_i|M_i) \doteq \beta^2((\sigma_{XB}^2\,\sigma_{YB}^2 + \sigma_{XYB}^2)/\sigma_{XY}^2 + 2\sigma_{XB}^4/\sigma_X^4 - 4\sigma_{XYB}\,\sigma_{XB}^2/\sigma_X^2\,\sigma_{XY}) + \sigma_Y^2K'/\sigma_X^2 \qquad (7.31)$$

where $K' = \rho^2(k_{22B}/\sigma_{XY}^2 + k_{4XB}/\sigma_X^4 - 2k_{31B}/\sigma_X^2\,\sigma_{XY})$

Since $\beta_B = \beta_W$ we have

$$\sigma_{XYB}^2/\sigma_{XY}^2 = \sigma_{XB}^4/\sigma_X^4 = \sigma_{XYB}\,\sigma_{XB}^2/\sigma_X^2\,\sigma_{XY}$$

Hence

$$\mathrm{var}_I(b_i|M_i) \doteq \beta^2(\sigma_{XB}^2\,\sigma_{YB}^2 - \sigma_{XYB}^2)/\sigma_{XY}^2 + \sigma_Y^2K'/\sigma_X^2$$

$$= \frac{\sigma_Y^2}{\sigma_X^2}\left[\tau_{Xm}\,\tau_{Ym}(1-\rho_B^2) + K'\right] \qquad (7.32)$$

Substituting into (6.41) and using Lemma 6.7 we have

$$\tilde{\tau}_{XYb} \doteq \frac{\tau_{Xm}\tau_{Ym}(1-\rho_B^2) + K'}{1-\rho^2 + K} \qquad (7.33)$$

If $K = K' = 0$ then

$$\tilde{\tau}_{XYb} \doteq \tau_{Xm}\tau_{Ym}(1-\rho_B^2)/(1-\rho^2)$$

$$= \tau_{Xm}\,\sigma_{Y|XB}^2/\,\sigma_{Y|X}^2$$

$$= \tau_{Xm}\tau_{Y|Xm} \quad , \text{ say.} \qquad (7.34)$$

where $\sigma_{Y|XB}^2 = \sigma_{YB}^2(1-\rho_B^2)$

$\sigma_{Y|X}^2 = \sigma_Y^2\,(1-\rho^2)$

$\tau_{Y|XM}$ may be interpreted as the residual intra-class correlation in Y not explained by the linear regression of Y on X.

Inference in regression analysis is conventionally performed conditionally on the x values, although, as Barndorff-Nielsen (1978, p.36) points out, the justification for this inferential separation is seldom given. The conditional distribution of the y values give the x values under Model I is given in Lemma 7.10. Normal within and between-cluster distributions are assumed since this gives linear regressions. Only the conditional distribution of $Y_i = (Y_{i1} \ldots Y_{imi})'$ given $X_i = (X_{i1} \ldots X_{imi})'$ is considered since the pairs (Y_i, X_i) are independent between clusters.

Lemma 7.10

Let $Y_i = (Y_{i1} \ldots Y_{im_i})'$, $X_i = (X_{i1} \ldots X_{im_i})'$

Suppose (i) Model I holds

(ii) Assumption A holds

(iii) $\sigma_{Xi}^2 = \sigma_{XW}^2$, $\sigma_{Yi}^2 = \sigma_{YW}^2$, $\sigma_{XYi} = \sigma_{XYW}$

(iv) $$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \Big| \begin{pmatrix} \mu_{Yi} \\ \mu_{Xi} \end{pmatrix} \sim N \left((I_2 \otimes 1_{m_i}) \begin{pmatrix} \mu_{Yi} \\ \mu_{Xi} \end{pmatrix}, \Sigma_W \otimes I_{m_i} \right) \quad (7.35)$$

and $$\begin{pmatrix} \mu_{Yi} \\ \mu_{Xi} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \Sigma_B \right) \quad (7.36)$$

where $$\Sigma_W = \begin{pmatrix} \sigma_{YW}^2 & \sigma_{XYW} \\ \sigma_{XYW} & \sigma_{XW}^2 \end{pmatrix} , \Sigma_B = \begin{pmatrix} \sigma_{YB}^2 & \sigma_{XYB} \\ \sigma_{XYB} & \sigma_{XB}^2 \end{pmatrix}$$

Then $$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N \left((I_2 \otimes 1_{m_i}) \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix} , \Sigma_W \otimes P_{Wi} + (m_i \Sigma_B + \Sigma_W) \otimes P_{Bi} \right) \quad (7.37)$$

$$Y_i|X_i \sim N \left(1_{m_i} \mu_{Y|Xi} + \beta_W P_{Wi} X_i + \bar{\beta}_i P_{Bi} X_i , \sigma_{Y|XW}^2 P_{Wi} + \overline{\sigma_{Y|Xi}^2} P_{Bi} \right) \quad (7.38)$$

where $\mu_{Y|Xi} = \mu_Y - \bar{\beta}_i \mu_X$

$$\bar{\beta}_i = (m_i\sigma_{XYB} + \sigma_{XYW})/(m_i\sigma_{XB}^2 + \sigma_{XW}^2)$$

$$P_{Bi} = 1_{m_i} 1_{m_i}'/m_i$$

$$P_{Wi} = I_{m_i} - P_{Bi}$$

$$\sigma_{Y|XW}^2 = \sigma_{YW}^2 - \sigma_{XYW}^2/\sigma_{XW}^2$$

$$\overline{\sigma_{Y|Xi}^2} = m_i\sigma_{YB}^2 + \sigma_{YW}^2 - (m_i\sigma_{XYB}^2 + \sigma_{XYW}^2)^2/(m_i\sigma_{XB}^2 + \sigma_{XW}^2)$$

Proof: Using standard results on normal mixtures of normal distributions (e.g. Lindley and Smith, 1972, pp.4, 5), it follows from (7.35) and (7.36) that

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N\left((I_2 \otimes 1_{m_i}) \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \Sigma_W \otimes I_{m_i} + (I_2 \otimes 1_{m_i}) \Sigma_B (I_2 \otimes 1_{m_i}') \right)$$

(7.37) follows immediately. (7.38) then follows using standard results on conditional distributions for multivariate normal random variables (e.g. Anderson, 1958 p.29).

Corollary 7.11

If, in addition to assumptions (i) - (iv) of Lemma 7.10 we assume

(v) $\beta_B = \beta_W = \beta$

$$Y_i|X_i \sim N(1_{m_i} \mu_{Y|X} + \beta X_i, \sigma_{Y|XW}^2 P_{Wi} + (m_i\sigma_{Y|XB}^2 + \sigma_{Y|XW}^2)P_{Bi}) \qquad (7.39)$$

where

$$\mu_{Y|X} = \mu_Y - \beta \mu_X$$

$$\sigma_{Y|XB}^2 = \sigma_{YB}^2 - \sigma_{XYB}^2/\sigma_{XB}^2$$

Proof: $\bar{\beta}_i = \lambda_i\beta_B + (1-\lambda_i)\beta_W$ where $\lambda_i = m_i\sigma_{XB}^2/(m_i\sigma_{XB}^2 + \sigma_{XW}^2)$

$= \beta$ if (v) holds

$$\overline{\sigma_{Y|Xi}}^2 = m_i\sigma_{YB}^2 + \sigma_{YW}^2 - \bar{\beta}_i^2(m_i\sigma_{XB}^2 + \sigma_{XW}^2)$$

$$= m_i\sigma_{Y|XB}^2 + \sigma_{Y|XW}^2$$

Lemma 7.12

If Model II holds and (X_{ij}, Y_{ij}) are jointly normally distributed (as in 7.32) then

$$Y_{ij}|X_{ij} \sim N(\mu_{Y|X} + \beta X_{ij} , \sigma_{Y|X}^2) \tag{7.40}$$

where $\sigma_{Y|X}^2 = \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2$

If $\beta_B = \beta_W$ then

$$\sigma_{Y|X}^2 = \sigma_{Y|XW}^2 + \sigma_{Y|XB}^2 \tag{7.41}$$

Proof: (7.40) is standard normal distribution theory. To obtain (7.41):

$$\begin{aligned}\sigma_{Y|X}^2 &= \sigma_Y^2 - \beta^2\sigma_X^2 \\ &= \sigma_{YW}^2 + \sigma_{YB}^2 - \beta^2(\sigma_{XW}^2 + \sigma_{XB}^2) \\ &= \sigma_{YW}^2 - \beta_W^2\sigma_{XW}^2 + \sigma_{YB}^2 - \beta_B^2\sigma_{XB}^2 \\ &\quad \text{if } \beta_B = \beta_W = \beta \\ &= \sigma_{Y|XW}^2 + \sigma_{Y|XB}^2\end{aligned}$$

It follows from Corollary 7.11 that when assumptions (i) - (v) hold (i.e. Case 1) the parameter space of Model I may be jointly indexed by the pairs of vectors $(\mu_{Y|X}, \beta, \sigma_{Y|XW}^2, \sigma_{Y|XB}^2)$ and $(\mu_X, \sigma_{XW}^2, \sigma_{XB}^2)$. The conditional distribution of the Y values given the X values only depends on $(\mu_{Y|X}, \beta, \sigma_{Y|XW}^2, \sigma_{Y|XB}^2)$ and the marginal distribution of the X values only depends on $(\mu_X, \sigma_{XW}^2, \sigma_{XB}^2)$. Hence from Definition 1.2 the X values are ancillary for $(\mu_{Y|X}, \beta, \sigma_{Y|XW}^2, \sigma_{Y|XB}^2)$ and, as argued in Section 1.2.3, it seems reasonable to make inference about β conditional on the X values. From Lemma 7.12 it also seems reasonable to use conditional inference for Model II.

In the following Lemma all moments are assumed to be conditional on s and $\underline{M}$.

Lemma 7.13

Under the assumptions of Corollary 7.11 for Model I and of Lemma 7.12 (with $\beta_B = \beta_W$) for Model II.

$$E_I(T_{XYb}|\underline{X}) = E_{II}(T_{XYb}|\underline{X}) = \beta$$

$$\text{var}_I(T_{XYb}|\underline{X})/\text{var}_{II}(T_{XYb}|\underline{X}) = 1 + (m^*-1)\tilde{\tau}_{Xm}\tau_{Y|Xm} \qquad (7.42)$$

where $\underline{X} = (X_{11} \ldots X_{nmn})$

$$\tilde{\tau}_{Xm} = \frac{\sum_i \sum_{j\neq K} (X_{ij}-T_{Xm})(X_{iK}-T_{Xm})}{\sum_i m_i(m_i-1) \sum_{ij} (X_{ij}-T_{Xm})^2 / \sum_i m_i}$$

and $\tau_{Y|Xm}$ is given in (7.34)

Proof: From (7.21) $T_{XYb} = \sum_i \sum_j w_{ij} Y_{ij}$

where $w_{ij} = (X_{ij}-T_{Xm})/\Sigma\Sigma(X_{ij}-T_{Xm})^2$

Hence $E_I(T_{XYb}|\underline{X}) = E_{II}(T_{XYb}|\underline{X})$ by definition of Model II

$$= \Sigma\Sigma w_{ij}(\mu_{Y|X} + \beta X_{ij}) \quad \text{from Corollary 7.11}$$

$$= \beta$$

since $\Sigma\Sigma w_{ij} = 0 \qquad \Sigma\Sigma w_{ij}X_{ij} = 1$

Let $w_i' = (w_{i1} \ldots w_{im_i})$

Then

$$T_{XYb} = \sum_i w_i' Y_i$$

Hence

$$\text{var}_I(T_{XYb}|\underline{X}) = \sum_i w_i' \, \text{var}_I(Y_i|\underline{X}) \, w_i$$

$$= \sum_i w_i'(\sigma^2_{Y|XW} P_{Wi} + (m_i\sigma^2_{Y|XB} + \sigma^2_{Y|XW})P_{Bi})w_i$$

from Corollary 7.11.

$$= \sigma_{Y|X}^2 \Sigma\Sigma w_{ij}^2 + \sigma_{Y|XB}^2 (\Sigma m_i w_i' P_{Bi} w_i - \Sigma\Sigma w_{ij}^2)$$

since $\sigma_{Y|X}^2 = \sigma_{Y|XW}^2 + \sigma_{Y|XB}^2$

$$= \sigma_{Y|X}^2 \Sigma\Sigma w_{ij}^2 + \sigma_{Y|XB}^2 \sum_i \sum_{j \neq k} (X_{ij} - T_{Xm})(X_{ik} - T_{Xm})$$

$$= \Sigma\Sigma w_{ij}^2 (\sigma_{Y|X}^2 + \sigma_{Y|XB}^2 (m^* - 1) \tilde{\tau}_{Xm}) \qquad (7.43)$$

From Lemma 7.12

$$\text{var}_{II}(T_{XYb} | \underline{X}) = \Sigma\Sigma w_{ij}^2 \sigma_{Y|X}^2 \qquad (7.44)$$

Dividing (7.43) by (7.44) gives (7.42).

Expression (7.42) is given by Campbell (1977). An alternative form for $\tilde{\tau}_{Xm}$ is given by Holt and Scott (1981) using the identity.

$$(m^* - 1)\tilde{\tau}_{Xm} = \frac{\Sigma m_i^2 (\bar{X}_i - T_{Xm})^2}{\Sigma\Sigma (X_{ij} - T_{Xm})^2} - 1$$

Note that expression (7.42) corresponds to the expression for $\tilde{\tau}_{XYb}$ in (7.34) where the superpopulation value of τ_{Xm} rather than the sample value is used.

The estimation of a difference between subclass means may be viewed as a special case of regression analysis where X only takes two values. The difference in means then corresponds to β. In this case $\tilde{\tau}_X$ can be interpreted as a measure of 'crossclassedness' (Kish et al, 1976). If $\tilde{\tau}_X = 1$ then the subclasses are completely 'segregated' and

$$\text{var}_I(T_{XYb} | \underline{X}) / \text{var}_{II}(T_{XYb} | \underline{X}) = 1 + (m^* - 1)\tau_{Y|Xm}$$

$$= \text{var}_I(T_{Ym} | \underline{X}) / \text{var}_{II}(T_{Ym} | \underline{X})$$

If $\tilde{\tau}_X = 0$ then the subclasses are completely 'crossed' with the clusters and $var_I(T_{XYb}|\underline{X})/var_{II}(T_{XYb}|\underline{X})$ attains its minimum. This is in accordance with the conventional wisdom on deffs for differences in subclass means (Kish et al, 1976).

Case 2 : A holds, $\sigma_{Xi}^2 = \sigma_{XW}^2$, $\sigma_{Yi}^2 = \sigma_{YW}^2$, $\sigma_{XYi} = \sigma_{XYW}$

β_B and β_W are now allowed to differ. (7.31) again holds and it may alternatively be written

$$var_I(b_i|M_i) \doteq \frac{\sigma_Y^2}{\sigma_X^2}\left[\tau_{Xm}\tau_{Ym}(1-\rho_B^2) + 2\tau_{Xm}^2\frac{\sigma_X^2}{\sigma_Y^2}(\beta_B-\beta)^2+K'\right] \qquad (7.45)$$

since $\frac{\sigma_Y^2}{\sigma_X^2}.2\tau_{Xm}^2\frac{\sigma_X^2}{\sigma_Y^2}(\beta_B-\beta)^2 = \frac{2\sigma_{XB}^4}{\sigma_X^4}\left(\frac{\sigma_{XYB}}{\sigma_{XB}^2}-\frac{\sigma_{XY}}{\sigma_X^2}\right)^2$

$$= \beta^2\left(\frac{2\sigma_{XYB}^2}{\sigma_{XY}^2}+\frac{2\sigma_{XB}^4}{\sigma_X^4}-\frac{4\sigma_{XYB}\sigma_{XB}^2}{\sigma_{XY}\sigma_X^2}\right)$$

Combining (7.26), (7.28) and (7.45) gives

$$\tilde{\tau}_{XYb} = \frac{\tau_{Xm}\,\tau_{Ym}(1-\rho_B^2) + 2\tau_{Xm}^2\sigma_X^2(\beta_B-\beta)^2/\sigma_Y^2 + K'}{1-\rho^2+K} \qquad (7.46)$$

Comparing (7.33) and (7.46) we see that the difference between β_B and β_W (recall from 7.30 that $\beta_B-\beta = (1-\tau_{Xm})(\beta_B-\beta_W)$) has inflated the misspecification effect of $\tilde{\tau}_{XYb}$.

Is it possible, in this case, to evaluate misspecification effects conditional on the X values as in Case 1? In general it seems inappropriate to do so. We argue this in two ways. (i) From Lemma 7.10 the conditional distribution of Y_i given X_i under Model I is indexed by the parameter vector $(\mu_{Y|Xi}, \beta_W, \bar{\beta}_i, \sigma^2_{Y|XW}, \bar{\sigma}^2_{Y|Xi})$. But

$$\beta = \psi_i\bar{\beta}_i + (1-\psi_i)\beta_W$$

where $\psi_i = (1 + (m_i-1)\tau_{Xm})/m_i$

and so β cannot be computed from the above parameter vector without knowledge of τ_{Xm} (unless $m_i=1$ or $\beta_B = \beta_W$). Hence it cannot be argued that the X values are ancillary for β under Model I in this case. (ii) From Lemma 7.10 we may express the conditional distribution of the Y values given the X values and the μ_{Xi} in terms of a linear model

$$Y_{ij} = \mu_Y + \beta_B(\mu_{Xi}-\mu_X) + \beta_W(X_{ij}-\mu_{Xi}) + \delta_i + \varepsilon_{ij} \qquad (7.47)$$

Now the μ_{Xi} are unobserved but we do know the sample cluster means $\bar{x}_i = \sum_i x_{ij}/m_i$ and we may write

$$\bar{x}_i = \mu_{Xi} + \eta_i \qquad (7.48)$$

(7.47) and (7.48) define a classical 'errors in variables regression' model for which the classical mode of analysis is unconditional on the X values. Holt and Scott (1981) replace μ_{Xi} by $\bar{x}_i$ in (7.47) (and hence essentially replace β_B by $\bar{\beta}_i$) and obtain expressions for the design effects of β_W and β_B.

Case 3 : A holds, $\mu_{Xi} = \mu_X$

In this case

$$b_i = \beta_i = \sigma_{XYi}/\sigma_{Xi}\,\sigma_{Yi}$$

and so from (7.26) and (7.28)

$$\tau_{XYb}^{\sim} \doteq \mathrm{var}_I(\beta_i)\ \sigma_X^2/\sigma_Y^2(1-\rho^2+K) \qquad (7.49)$$

Hence even when τ_{Xm} and possibly τ_{Ym} are zero $\tau_{XYb}^{\sim}$ need not be zero. It is interesting that $\tau_{XYb}^{\sim}$ does not depend on the residual intra-cluster correlation $\tau_{Y|Xm}$. To understand this we note that the between-cluster component of (7.47) drops out and we may write

$$Y_{ij} = \mu_Y + \beta_i(X_{ij}-\mu_X) + \delta_i + \varepsilon_{ij} \qquad (7.50)$$

Now $T_{XYb} \propto \sum_i \sum_j (X_{ij}-T_{Xm})Y_{ij}$

$$= \sum_i \sum_j (X_{ij}-T_{Xm})(\beta_i(X_{ij}-\mu_X) + \delta_i + \varepsilon_{ij}) \qquad (7.51)$$

But if $\mu_{Xi} = \mu_X$ then we would expect $\sum_j (X_{ij}-T_{Xm})$ to be small and so (7.51) is approximately

$$\sum_i \beta_i \sum_j (X_{ij}-T_{Xm})^2 + \sum_i \sum_j (X_{ij}-T_{Xm})\varepsilon_{ij}$$

The δ_i term, which induces the residual intra-cluster correlation $\tau_{Y|Xm}$, is negligible.

Case 4 : A holds

As for $\tau_{XYc}^{\sim}$ and $\tau_{XYr}^{\sim}$ we may write

$$\tau_{XYb}^{\sim} = \tau_1 + \tau_2 + \tau_3$$

where τ_1 is given by (7.33), τ_3 by (7.49) and τ_2 is such that

$$|\tau_2| \leq 2(\tau_1\ \tau_3)^{\frac{1}{2}}$$

Some estimates of these quantities, under the assumption that X_{ij} and Y_{ij} are normally distributed within clusters are given in Table 7.2.

Table 7.2	Estimates for FES Data			
Variables X,Y	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_{XYb}^{\sim}$
log(V1),log(V2)	.0005	-.0012	.0433	.0426
log(V1),log(V3)	.0002	-.0005	.0433	.0431
log(V3),log(V2)	.0006	-.0004	.0295	.0298

Again $\hat{\tau}_{XYb}^{\sim}$ is dominated by $\hat{\tau}_3$ which is proportional to the variation in cluster regression coefficients, β_i. The values of $\tau_{XYb}^{\sim}$ are

smaller than those for $\tilde{\tau}_{XYr}$ in Table 7.1 but still greater than the corresponding values for $\hat{\tau}_{Xm}$.

Case 5 : B holds but A does not hold

By analogy with the results of $\tilde{\tau}_{XYr}$ we may write

$$\tilde{\tau}_{XYb}(M_i) = \int_0^1 \left[r_{XY}^2(\alpha_i t) + r_X(\alpha_i t)r_Y(\alpha_i t) + 2\rho^2 r_X^2(\alpha_i t) - 4\rho r_X(\alpha_i t)r_{XY}(\alpha_i t)\right] K(\alpha_i t)dt/(1-\rho^2)$$

Again expression for $\tilde{\tau}_{XYb}$ as a function of M_i may be obtained by substituting particular functional forms for K.

7.3 Principal Components Analysis

As in Section 4.3, we restrict attention to the eigenvalues $\ell_1 \geq \ldots \geq \ell_p$ and corresponding eigenvectors $g_1 \ldots g_p$ of the sample covariance matrix,

$$S_s = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij}-\bar{y}_s)(y_{ij}-\bar{y}_s)'/(m_o-1) \qquad (7.52)$$

The first and second moments of the ℓ_k and g_k may be obtained by substituting the results of Sections 5.4 and 5.5 into Corollary 4.8. These results will be approximate for large n . Under Assumption B the ℓ_k and g_k will be approximately unbiased for the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ and the corresponding eigenvectors $\gamma_1 \ldots \gamma_p$ respectively of the super population covariance matrix Σ under Model I. Now ℓ_k is the sample variance of the variate $g_k'y_{ij} = \gamma_k'y_{ij} + O_p(n^{-\frac{1}{2}})$ and the variate $\gamma_k'y_{ij}$ has variance $\gamma_i'\Sigma\gamma_k = \lambda_k$. Hence the second moments of the ℓ_k may be obtained from Section 5.4. The second moment of the g_k are less easily derived. Rather than consider the full generality of Model I we restrict our investigation to the conventional multivariate one-way random effects model:

$$Y_{ij}|\theta_i \sim N_p(\mu_i, \Sigma_W) \qquad (7.53)$$

$$\mu_i \sim N_p(\mu, \Sigma_B)$$

We require the following result:

Lemma 7.14

Given (7.53)

$$\mathrm{cov}_I(S_{s\alpha\beta}, S_{s\gamma\delta} | s, \underline{M}) = (\Sigma_{\alpha\gamma}\Sigma_{\beta\delta}+\Sigma_{\alpha\delta}\Sigma_{\beta\gamma})/m_o + (m^*-1)(\Sigma_{B\alpha\gamma}\Sigma_{B\beta\delta} + \Sigma_{B\alpha\delta}\Sigma_{B\beta\gamma})/m_o$$

Proof:

As in Sections 5.4 and 5.5 we may approximate S_s to order n^{-1} by

$$\tilde{S_s} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij}-\mu)(y_{ij}-\mu)'/m_o$$

Hence

$$\mathrm{cov}_I(S_{s\alpha\beta}, S_{s\gamma\delta} | s, \underline{M}) \doteq E_I\left[\mathrm{cov}_I(\tilde{S}_{s\alpha\beta}, \tilde{S}_{s\gamma\delta} | \theta_i, s, \underline{M}) | s, \underline{M}\right]$$

$$+ \mathrm{cov}_I\left[E_I(\tilde{S}_{s\alpha\beta} | \theta_i, s, \underline{M}), E_I(\tilde{S}_{s\gamma\delta} | \theta_i, s, \underline{M}) | s, \underline{M}\right]$$

$$= E_I\Big[\Sigma_{W\alpha\gamma}\Sigma_{W\beta\delta} + \Sigma_{W\alpha\delta}\Sigma_{W\beta\gamma} + (\mu_{i\alpha}-\mu_\alpha)(\mu_{i\gamma}-\mu_\gamma)\Sigma_{W\beta\delta}$$

$$+ (\mu_{i\beta}-\mu_\beta)(\mu_{i\gamma}-\mu_\gamma)\Sigma_{W\alpha\delta} + (\mu_{i\alpha}-\mu_\alpha)(\mu_{i\delta}-\mu_\delta)\Sigma_{W\beta\gamma}$$

$$+ (\mu_{i\beta}-\mu_\beta)(\mu_{i\delta}-\mu_\delta)\Sigma_{W\alpha\gamma} | s, \underline{M}\Big]/m_o$$

$$+ \mathrm{cov}_I\Big[\Sigma m_i(\mu_{i\alpha}-\mu_\alpha)(\mu_{i\beta}-\mu_\beta),\ \Sigma m_j(\mu_{j\gamma}-\mu_\gamma)(\mu_{j\delta}-\mu_\delta) | s, \underline{M}\Big]/m_o^2$$

$$= \Big[\Sigma_{W\alpha\gamma}\Sigma_{W\beta\delta} + \Sigma_{W\alpha\delta}\Sigma_{W\beta\gamma} + \Sigma_{B\alpha\gamma}\Sigma_{W\beta\delta} + \Sigma_{B\beta\gamma}\Sigma_{W\alpha\delta}$$

$$+ \Sigma_{B\alpha\delta}\Sigma_{W\beta\gamma}\Big]/m_o + \Sigma m_i^2(\Sigma_{B\alpha\gamma}\Sigma_{B\beta\delta} + \Sigma_{B\alpha\delta}\Sigma_{B\beta\gamma})/m_o^2$$

$$= (\Sigma_{\alpha\gamma}\Sigma_{\beta\delta} + \Sigma_{\alpha\delta}\Sigma_{\beta\gamma})/m_o + (m^*-1)(\Sigma_{B\alpha\gamma}\Sigma_{B\beta\delta}+\Sigma_{B\alpha\delta}\Sigma_{B\beta\gamma})/m_o$$

since

$$\Sigma = \Sigma_B + \Sigma_W$$

We may now obtain the approximate first two moments of the eigenvalues of S_s

Lemma 7.15

Given (7.53)

$$E_I(\ell_k|s,\underline{M}) \doteq \lambda_k$$

$$V_I(\ell_k|s,\underline{M}) \doteq (1 + (m^*-1)\tau^2_{\lambda k})2\lambda^2_k/m_o$$

where

$$\tau_{\lambda k} = \gamma_k'\Sigma_B\gamma_k/\lambda_k$$

$$= V_I(\gamma_k'\mu_i)/V_I(\gamma_k'Y_{ij})$$

Proof:

We have already noted that ℓ_k is approximately unbiased for λ_k for large n.

From Corollary 4.8:

$$V_I(\ell_k|s,\underline{M}) \doteq \sum_{\alpha\beta\delta\gamma} (\gamma_k)_\alpha(\gamma_k)_\beta(\gamma_k)_\delta(\gamma_k)_\gamma \, cov_I(S_{s\alpha\beta}, S_{s\delta\gamma}|s,\underline{M})$$

$$\doteq 2(\gamma_k'\Sigma\gamma_k)^2/m_o + 2(m^*-1)(\gamma_k'\Sigma_B\gamma_k)^2/m_o$$

from Lemma 7.14

$$= 2\lambda^2_k/m_o\left[1 + (m^*-1)\tau^2_{\lambda k}\right]$$

as required.

As remarked before, this result could have been obtained directly by substituting the variate $\gamma_k'Y_{ij}$ into Case 1 of Section 5.4 (see 5.51), where the misspecification effect of a variance $T_{Y\tilde{v}}$ was obtained as $1 + (m^*-1)\tau^2_{Ym}$ and $\tau_{Ym} = \sigma^2_B/\sigma^2$. The broad interpretation of Lemma 7.15 is that the variance of ℓ_k will be largest when the 'direction of largest variation' of the μ_i is along γ_k. At one extreme if the γ_i are confined to a hyperplane orthogonal to γ_k

then $\tau_{\lambda k} = 0$ and at the other extreme if the μ_i are confined to a one-dimensional line in the direction of γ_k then $\Sigma_B = \alpha\lambda_k\gamma_k\gamma_k'$ where $0 \leq \alpha \leq 1$ and $\tau_{\lambda k} = \alpha$.

Lemma 7.16

Given (7.53)

$$E_I(g_k|s,\underline{M}) \doteq \gamma_k$$

$$V_I(g_k|s,\underline{M}) \doteq \sum_{j\neq k} \lambda_k\lambda_j\gamma_j\gamma_j'/m_o(\lambda_j-\lambda_k)^2 + (m^*-1) \sum_{i\neq k}\sum_{j\neq k}(\gamma_k'\Sigma_B\gamma_j\gamma_i'\Sigma_B\gamma_k$$

$$+ \gamma_k'\Sigma_B\gamma_k\gamma_i'\Sigma_B\gamma_j)\gamma_i\gamma_j'/(\lambda_k-\lambda_i)(\lambda_k-\lambda_j)m_o \qquad (7.54)$$

Proof:

From Corollary 4.8

$$V_I(g_k|s,\underline{M}) = \sum_{i\neq k}\sum_{j\neq k} \mathrm{cov}(w_{ki},w_{kj})\gamma_i\gamma_j'$$

where

$$\mathrm{cov}(w_{ki},w_{kj}) = \sum_{\alpha\beta\delta\gamma} (\gamma_k)_\alpha(\gamma_i)_\beta(\gamma_k)_\delta(\gamma_j)_\gamma \mathrm{cov}(S_{\alpha\beta},S_{\gamma\delta})/(\lambda_k-\lambda_i)(\lambda_k-\lambda_j)$$

$$= \Big[(\gamma_k'\Sigma\gamma_j)(\gamma_i'\Sigma\gamma_k) + (\gamma_k'\Sigma\gamma_k\gamma_i'\Sigma\gamma_j)+(m^*-1)(\gamma_k'\Sigma_B\gamma_j\gamma_i'\Sigma_B\gamma_k$$

$$+ \gamma_k'\Sigma_B\gamma_k\gamma_i'\Sigma_B\gamma_j)\Big]/(\lambda_k-\lambda_i)(\lambda_k-\lambda_j)m_o$$

$$= \Big[\lambda_k\lambda_i\delta_{ij} + (m^*-1)(\gamma_k'\Sigma_B\gamma_j\gamma_i'\Sigma_B\gamma_k + \gamma_k'\Sigma_B\gamma_k\gamma_i'\Sigma_B\gamma_j)\Big]/(\lambda_k-\lambda_i)(\lambda_k-\lambda_j)m_o$$

and the result follows

Special Case 1 : $\Sigma_B = \alpha_\ell\lambda_\ell\gamma_\ell\gamma_\ell'$ $\qquad 0 \leq \alpha_\ell \leq 1$

In this case the second term in (7.54) disappears for all k and so there is no increase in imprecision in g_k over the IID Model II case (see Corollary 4.9).

Special Case 2 : $\Sigma_B = \alpha_\ell \lambda_\ell \gamma_\ell \gamma_\ell' + \alpha_m \lambda_m \gamma_m \gamma_m' \qquad 0 \leq \alpha_\ell, \alpha_m \leq 1$

If k is not equal to ℓ or m then the second term in (7.54) disappears and there is no increase in precision. If $k = \ell$, say then

$$V_I(g_k|s,\underline{M}) = \sum_{j \neq km} \lambda_k \lambda_j \gamma_j \gamma_j' / m_o(\lambda_j - \lambda_k)^2 + \left(1+(m^*-1)\alpha_k\alpha_m\right)\lambda_k \lambda_m \gamma_m \gamma_m' / m_o(\lambda_m - \lambda_k)^2$$

Hence the 'intra-cluster correlation' enters again via a product $\alpha_k \alpha_m$. The misspecification effect will clearly be largest when λ_m is close to λ_k. It seems fair to attempt a broad generalisation. g_k will be most unstable when both (1) the μ_i vary in the direction of γ_k and (2) the μ_i vary in the direction of the γ_m for which $\lambda_m - \lambda_k$ is small.

7.4 Factor Analysis

In the standard approach to factor analysis we assume (c.f Section 4.4)

$$Y_{ij} = \mu + \Lambda f_{ij} + u_{ij} \qquad (7.55)$$

where

$$f_{ij} \sim \text{NID}_m(0,\Phi) \ , \ u_{ij} \sim \text{NID}_p(0,\Psi) \ ,$$

the f_{ij} and u_{ij} are independent, Ψ is diagonal and the parameters are $(\mu, \Lambda, \Phi, \Psi)$.

The parameters may be estimated by maximum likelihood (e.g. Lawley and Maxwell, 1971). These ML estimators are a function of the sample covariance matrix S_s (of 7.52) and so, in principle, we could obtain the distribution of these estimators given the alternative distribution of S_s under Model I, e.g. using the results of Fuller et al. (1982). Such an approach only seems useful for variance estimation, however, and offers little theoretical insight. Instead we consider simplifying Model I.

Beginning with the linear model formulation (7.55) we might conceive of an intra-cluster dependence of the f_{ij} and u_{ij}. Let us consider an example due to Muthén (1981). The Y variables are responses to questions about attitudes to abortion. Muthén fits a two factor solution (m = 2). The first factor, labelled 'Medical', distinguishes between those people who find medical factors to be good reasons for justifying abortion and those who do not. The second factor, labelled 'social', distinguishes between those who find social factors to be good reasons and those who do not. In subsequent analysis Muthén finds that individuals' scores on the factors differ particularly between religious groups (Protestant and Catholic) and to a lesser extent between individuals with different levels of education. In a similar manner it would seem likely that a clustered survey of the U.K. would exhibit intra-cluster correlation on these common factors e.g. we would expect very different factor scores in a working class area of Belfast compared with a middle-class area of London. On the other hand, as was argued in support of Meredith's (1964) model in Section 4.4, it is not obvious that there should be intracluster correlation on the unique factors u_{ij}. If the u_{ij} represent measurement error and 'non-behavioural' unique components then these may not be associated with the socio-economic factors underlying the clustering. As noted in Section 4.4, the most likely effect of clustering on the u_{ij} would be for the Ψ matrix to differ between clusters. Let us, however, only consider the simplest possible model with intra-cluster correlation for the f_{ij}.

$$Y_{ij} = \mu + \Lambda f_{ij} + u_{ij}$$

$$f_{ij} = \alpha_i + \varepsilon_{ij}$$

$$u_{ij} \sim N_p(0,\Psi),\ \alpha_i \sim N_m(0,\Phi_B),\ \varepsilon_{ij} \sim N_m(0,\Phi_W)$$

where the u_{ij}, α_i and ε_{ij} are mutually independent. Note that

$$V(Y_{ij}) = \Sigma = \Lambda\Phi\Lambda' + \Psi$$

$$V(f_{ij}) = \Phi = \Phi_B + \Phi_W$$

Now the sample covariance matrix, S_s (see 7.52), of the y_{ij} may be written

$$S_s = (S_B + S_W)/(m_o-1)$$

where

$$S_B = \sum_{i=1}^{n} m_i(\bar{y}_i-\bar{y}_s)(\bar{y}_i-\bar{y}_s)'$$

$$S_W = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij}-\bar{y}_i)(y_{ij}-\bar{y}_i)'$$

In the simplest case of $m_i = m \quad i = 1...n$, S_B and S_W are independent and

$$S_B \sim W_p(n-1, \Lambda\Phi_W\Lambda' + \Psi + m\Lambda\Phi_B\Lambda')$$

$$S_W \sim W_p(n(m-1), \Lambda\Phi_W\Lambda' + \Psi)$$

Maximum likelihood estimates of Λ, Ψ, Φ_B and Φ_W may then be obtained by using the 'simultaneous factor analysis for several populations' option in LISREL (Joreskog and Sorbom, 1978). S_B would be treated as the sample covariance matrix of a random sample of n observations from one population and S_W as the sample covariance matrix of a random sample of $n(m-1) + 1$ observations from a second population. Using LISREL the factor loading matrix, Λ, and the specific variance matrix, Ψ, may be constrained to be equal in both populations. The only problem might be that the estimate $\hat{V}$ of the covariance matrix, $V = \Phi_W + m\Phi_B$, in the first population and the estimate $\hat{\Phi}_W$ from the second population were such that $\hat{V} - \hat{\Phi}_W = m\hat{\Phi}_B$ was not positive semi-definite. This should not be disturbing if one is only interested in the estimate of Φ, $\hat{\Phi} = \hat{V}/m + (m-1)\hat{\Phi}_W/m$, which will be positive semi-definite, providing $\hat{V}$ and $\hat{\Phi}_W$ are.

We have now suggested 'alternative' estimators of the parameters under our simplified model. The question of the properties of the standard estimators remains. The simplification of the model has not made this problem essentially easier because, although the distribution of S_s is more straightforward, the standard estimators are still very complicated

functions of S_s . For this reasons we only attempt a heuristic approach. As n increases

$$S_B/n \xrightarrow{P} \Lambda\phi_W\Lambda' + \Psi + m\Lambda\Phi\Lambda'$$

$$S_W/n \xrightarrow{P} (m-1)(\Lambda\Phi_W\Lambda'+\Psi)$$

and so

$$S_s \xrightarrow{P} \left(\Lambda\phi_W\Lambda'+\Psi+m\Lambda\Phi\Lambda'+(m-1)(\Lambda\Phi_W\Lambda'+\Psi)\right)/m = \Lambda\Phi\Lambda'+\Psi$$

and so the estimators will be consistent. The problem occurs when n is fixed and m increases. In this case S_W/m will converge to $n(\Lambda\Phi_W\Lambda'+\Psi)$ but S_B/m will not converge. Rather it will act as $n\Lambda S_{\alpha B}\Lambda'$ where $S_{\alpha B}$ is the between group covariance matrix of the α_i

$$S_{\alpha B} = \sum_{i=1}^{n} (\alpha_i-\bar{\alpha})(\alpha_1-\bar{\alpha})'/n$$

Hence S_s will approach

$$\Lambda(\Phi_W+S_{\alpha B})\Lambda' + \Psi$$

For this reason we conjecture that intra-cluster correlation in our simplified model should have little effect on the estimates of Ψ and Λ, subject to rotation, but that it will inflate the variance of the estimates of Φ. We are hesitant in making this conjecture for at least two reasons. Firstly our argument depends on m being large, an assumption we have not previously used. Secondly the qualification 'subject to rotation' is very vague. If we initially estimate an orthogonal factor model then the instability in $\hat{\Phi}$ will presumably affect $\hat{\Lambda}$. One can only conjecture that the instability in $\hat{\phi}$ would only perturb $\hat{\Lambda}$ within the column space of Λ , e.g. if $m = 1$ then it would only affect $\hat{\Lambda}$ proportionately.

CHAPTER EIGHT - CONCLUSION

8.1 Summary of Thesis

This theses began with the observation that many statistical methods, including multivariate methods, are frequently used with sample survey data without any consideration being given to the sample selection scheme. This raises two broad questions: (A) to what extent do such methods remain valid under various selection schemes (*robustness*) and (B) what alternative methods might be adopted (*optimality*)? We distinguished between two approaches. In the first (*dissaggregated*) approach it is argued that the purpose of the multivariate analysis of sample survey data is to explore and model the structure of the data in relation to the structure (e.g. stratification or clustering) of the population used in the sample design. According to this approach, question (A) may be answered negatively a priori since, by definition, standard methods take no account of such population structure. In the second (*aggregated*) approach it is argued that the population structure used in sample design is a priori irrelevant to the substantive questions of interest and hence any targets of inference should be characteristics of the finite population (or of an aggregate superpopulation model of which the finite population is a realisation). Whilst noting that the first approach might often be appropriate, we restricted the ambit of this thesis to the second aggregated approach. For this purpose both questions (A) and (B) are relevant. It should be noted, however, that we have characterised the aggregated and disaggregated approaches in a rather extreme manner. Substantial overlap between the approaches exists. For example, disaggregated within-stratum analyses may be viewed as several aggregated analyses by redefining the strata as populations.

More formally we distinguished between design-based and model-based approaches to statistical inference. This raises foundational questions which we have not addressed. For inference about finite population characteristics, the design-based approach does offer a practical and robust (in the sense of making few assumptions) procedure for answering question (B) and to a more limited extent answering

question (A). On the other hand the model-based approach offers more theoretical insight into both questions and extends naturally to a disaggregated approach which the design-based approach does not. Again, whilst recognising the possibilities of the design-based approach we have largely restricted the ambit of this thesis to the model-based approach.

We also compared the problems of making inference about finite population and superpopulation parameters. The latter target of inference seemed to us to be most natural in multivariate analysis but we have also, to a lesser extent, considered finite population parameters since this enables us to compare results with the design-based approach and also allows us to make fewer model assumptions (because a marginal distribution for the design variables need not be specified). In both cases we have limited our investigations to point estimation (and prediction).

This thesis divides into two distinct parts. In Part I (Chapters 2-4) we consider a selection scheme which is conceptually very general. The associated model is, however, like most classical multivariate analysis, rather restrictive assuming independent individual values and either multivariate normality of else a linear homoskedastic relationship between the survey variables and the design variables. As such, this framework is very convenient for assessing the effects of selection on classical multivariate methods (question A) without introducing extra distributional complications. The use of this model (and selection scheme) in deriving alternative estimators (question B) should be treated with more caution since the assumptions are so restrictive. In Chapter 2 we show how selection can induce bias into standard estimators of a mean vector or covariance matrix. In Chapter 3 we show how such bias can be overcome by means of regression-type estimators. In Chapter 4 we demonstrate similarly the existence of bias in standard estimators of correlation coefficients and regression coefficients and in principal components analysis and factor analysis. Again we suggest alternative estimators.

In Part II (Chapters 5-7) we consider a more specific and conventional

sampling scheme: two-stage sampling. We develop a model which is very general and define parameters of interest in terms of an aggregate superpopulation. In Chapter 5 we show that if the marginal distribution of values does not depend on the cluster sizes then model-based inference may be formally justified and the standard estimator are approximately model-unbiased. The variances of standard estimators are, however, inflated by a factor which generalises the conventional $1 + (m-1)\rho$ expression. If, on the other hand, the marginal distribution of values does depend on the cluster sizes then, provided the design is self-weighting, standard estimators appear to be approximately design-model unbiased and the design-model variance is inflated by a similar factor. The form of the inflation factor is investigated in detail for a number of statistics and model assumptions. In Chapter 6 alternative estimators of the aggregate superpopulation parameters and predictors of the finite population moments are considered. In Chapter 7 the theory of Chapters 5 and 6 is extended to the problems of estimating correlation coefficients and regression coefficients and to principal components analysis and factor analysis.

8.2 Conclusions and Suggestions for Further Work

Analytical surveys differ from descriptive surveys in a number of respects, in particular because the target of inference is seldom clear. It may be conjectured (see Chapter 1) that the problem of statistical inference also differs, at least in magnitude, because survey sampling designs have less impact on analytical inference than on descriptive inference. This might be argued from an a priori viewpoint, as one might attempt to justify haphazard sampling of subjects for psychophysiological experiments.

For example, in regression analysis we might suppose that an i^{th} individual's score y_i is determined by $y_i = \beta x_i + e_i$ where β is a scientific behavioural constant homogeneous across the population and the e_i are disturbances with distribution (given x_i) homogenous across the population. In the ideal case where the e_i represent 'pure' measurement error or 'Popperian' random behaviour and are independent of any selection variables, inference about β may proceed in an

identical manner however the sample is selected. In the more practical case where the e_i are related to the selection variables (e.g. they display intra-cluster correlation or systematic differences between strata) the impact of selection on inference about β will depend on the effect of selection on the distribution of the e_i. The impact of selection on inference about the population mean of the y_i's should be greater, however, being a combination of the effect of selection on the e_i *and* the effect of selection on the x_i.

This argument seems to break down in practice because there will also be an effect of selection on the regression relationship, for example there will be differences between strata or clusters (see Chapter 5). For this reason we have treated the problem of multivariate analysis as one of estimating the parameters of certain distributions across the population, a parallel inferential problem to that of descriptive surveys. As such we have avoided invoking deeper ideas of scientific 'structural models' (Koopmans, 1947).

In this thesis we have considered the choice and properties of point estimators under various models/sampling designs. In practice we may divide the sampling designs that we have considered into three categories.

(a) Stratified sampling (not based on auxiliary variable known for population).

In this case standard point estimates weighted by the inverses of the stratum sampling fractions have reasonable design-based and model-based (see Section 2.3) interpretations when the target of inference is an aggregate parameter.

(b) Sampling with auxiliary information

Suppose the values of an auxiliary variable (or variables), X, are known for all finite population units (or in certain circumstances only certain summary statistics for the finite population units will be known. We shall distinguish between two cases.

(i) X not used in selection

Here we may have srs with auxiliary information and the regression-type estimators of Chapter 3 will have improved (large sample) efficiency over the sample mean both from a design-based and a model-based point of view.

(ii) X used in selection

If X is used in a stratified or ppx design then the regression-type estimators of Chapter 3 will be model-unbiased and will generally differ from the conventional design-based estimators. If X is used for truncated sampling or is used as a proxy for non-response (e.g. Nathan, 1982) then no design-unbiased estimator will exist but the same regression-type estimators may be used from a model-based point of view.

(c) Multi-stage sampling

In this situation the choice of point estimators is much less clear-cut. Under Assumptions A or B of Section 5.1. the most natural design based estimators would be the ratio-type estimators (for srs) or the unweighted estimators (for pps) discussed in Section 6.4. Under the same assumptions either of these estimators or the expansion-type estimators discussed in Chapter 5 should have model-biases of small order and the choice of most efficient estimators will depend on the distributional properties of the model (Section 6.2). Further work is needed to investigate the choice of estimator when Assumption A or B of Section 5.1 does not hold.

In the remainder of this chapter we consider areas for further research. Since this thesis only contains preliminary results on the particular problem of point estimation, the most important direction to follow would be towards a more practical 'package' of statistical methods for the analysis of multivariate survey data. We break down areas for further work according to the division of this thesis.

1. Pearson-type Selection Scheme

(a) In Sections 3.2 and 3.3 we presented some regression-type estimators based on the assumption of multivariate normality. The properties of these estimators may be compared with the more general regression-type

estimators of Fuller (1982). Small sample properties of these estimators could be evaluated by Monte Carlo methods as in Holt et al (1980b) with special consideration being given to (i) robustness to departures from model assumptions and (ii) the measure of 'goodness' of the estimators e.g. with respect to the model or randomisation distribution.

(b) A practical advantage of the conventional regression estimator is that it provides constant weights (for given auxiliary variables) to apply to all survey variables. Possible analogous weights for the regression estimation of a covariance matrix should be investigated.

(c) Diagnostic tests for the use of the regression type estimators might be developed and the use of a procedure involving a preliminary test of significance might be evaluated using the approach of Grimes and Sukhatme (1980).

(d) The alternative estimation procedures used for factor analysis in Section 4.4 should be further investigated.

(e) Estimators of standard errors and interval estimates based on the regression-type estimator should be considered.

(f) Methods of hypothesis testing (e.g. F-tests in regression and LR tests in factor analysis) should be considered.

(g) The application of these methods to non-response problems and to sample selection problems considered by econometricians (e.g. Heckman, 1979) should be investigated as in the work of Nathan (1982).

(h) The problem of 'optimal' sample design for estimating covariances might be considered (c.f. Sedransk 1965a).

Two-stage Sampling

(a) In Chapter 6 we considered optimal model-based estimation and obtained only preliminary results of limited practical value. Simpler estimators should be developed for use under different model assumptions.

Generalised least squares estimation might provide the most straightforward approach. Diagnostic checks are needed to distinguish between different models. In particular the effect of dependence of the values on the cluster sizes needs more research. The problem of informative design can arise here. The relation between model-based and design-based estimators needs to be considered also.

(b) Estimates of standard errors and interval estimates are very important. The use of Lemma 5.10 for obtaining variance estimates without using partial derivatives should be investigated. Again comparisons should be made with design-based approaches, the properties of which might be evaluated under a model as in Fuller (1975).

(c) In practice clusters are usually nested within strata. Methods should be extended to this situation and also to multi-stage designs.

(d) The impact of clustering on hypothesis testing such as F-tests in regression (see Shah et al, 1977) should be considered.

(e) The estimation procedures might be extended to disaggregated modelling.

(f) Problems of sample design might be considered with special reference to the spatial process approach of Chapter 5.

Finally some case studies are needed. These might help not only to throw into perspective the relative practical importance of some of the issues discussed above, but also to make possible a comparison between the impact of survey design and other inferential problems. For example, in fitting regression models to Family Expenditure Survey data, Mkai (1981) shows that heteroskedasticity may have far greater effect on the standard errors of least squares estimators than does survey design. It would also be useful to develop an integrated approach to handling not only the impact of survey design but also that of measurement errors, another major problem for inference in analytical surveys (Fuller, 1975).

APPENDIX - PARAMETER ESTIMATION FOR TWO-STAGE MODEL

In this section we explain how the estimates were obtained for Tables 5.2-5.4, 7.1 and 7.2. We also consider how misspecification effects might be estimated. We assume that Model I of Section 5.1 is true and in addition that

(i) assumption A holds (A.1)

and (ii) the within-cluster distributions are normal and the between-cluster distribution of (μ_{Xi}, μ_{Yi}) is normal. (A.2)

These assumptions are strong but should not prevent us obtaining a rough idea of the relative orders of magnitude of the various components of the misspecification effects in Chapters 5 and 7 for the FES data. Under assumptions (A.1) and (A.2), we shall now show that all the parameters of interest may be expressed as functions of the four quantities below, where the notation is as in Chapters 5 and 7.

(i) σ_{XYW}

(ii) σ_{XYB}

(iii) $P(X, Y, Z, V) = \text{cov}_I(\sigma_{XYi}, \sigma_{ZVi})$

(iv) $Q(X, Y, Z, V) = \text{cov}_I(\sigma_{XYi}, (\mu_{Zi} - \mu_Z)(\mu_{Vi} - \mu_V))$

Table 5.2: Under assumptions (A.1) and (A.2), τ_{Ym}, w_{v1}, τ_{Yv1}, r_v, w_{v2}, τ_{Yv2}, $\tau_{Yv}^{\sim}$, $\text{meff}(T_{Ym})$, $\text{meff}(T_{Yv}^{\sim})$ are known functions of σ^2_{YB}, σ^2_{YW}, γ_Y and c_{1Y} since

$$\sigma^2_Y = \sigma^2_{YB} + \sigma^2_{YW}$$

$$k_{4Y} = 3\gamma_Y + 6c_{1Y}$$

$$\sigma^2_{YB} = \sigma_{YYB}, \quad \sigma^2_{YW} = \sigma_{YYW}$$

$$\gamma_Y = P(Y,Y,Y,Y), \quad c_{1Y} = Q(Y,Y,Y,Y)$$

Table 5.4: Under assumptions (A.1) and (A.2), τ_1, τ_2, τ_3 and τ_{XYc} are known functions of σ^2_{XB}, σ^2_{XW}, σ^2_{YB}, σ^2_{YW}, σ_{XYB}, σ_{XYW}, c_{1XY}, γ_{XY}, $c_{X\cdot Y}$ $c_{Y\cdot X}$, $\delta_{X\cdot Y}$ since

$$\rho_B = \sigma_{XYB}/\sigma_{XB}\,\sigma_{YB}$$

$$\rho = (\sigma_{XYB} + \sigma_{XYW})/\sigma_X\sigma_Y$$

$$k_{22} = 4c_{1XY} + c_{X\cdot Y} + c_{Y\cdot X} + 2\gamma_{XY} + \delta_{X\cdot Y}$$

And $\sigma^2_{XB} = \sigma_{XXB}$, $\sigma^2_{XW} = \sigma_{XXW}$

$$c_{1XY} = Q(X,Y,X,Y),\ \gamma_{XY} = P(X,Y,X,Y)$$

$$c_{X\cdot Y} = Q(X,X,Y,Y),\ c_{Y\cdot X} = Q(Y,Y,X,X)$$

$$\delta_{X\cdot Y} = P(X,X,Y,Y)$$

Table 7.1: Under assumptions (A.1) and (A.2), τ_1, τ_2, τ_3 and τ_{XYr} are known functions of σ^2_{XB}, σ^2_{XW}, σ^2_{YB}, σ^2_{YW}, σ_{XYB}, σ_{XYW}, c_{1X}, c_{1Y}, c_{1XY}, γ_X, γ_Y, γ_{XY}, $c_{X\cdot Y}$, $c_{Y\cdot X}$, $\delta_{X\cdot Y}$, $c_{XY\cdot X}$, $c_{X\cdot XY}$, $c_{XY\cdot Y}$, $c_{Y\cdot XY}$, $\delta_{X\cdot XY}$, $\delta_{Y\cdot XY}$ since

$$k_{13} = c_{XY\cdot X} + c_{X\cdot XY} + \delta_{X\cdot XY}$$

$$k_{31} = c_{XY\cdot Y} + c_{Y\cdot XY} + \delta_{Y\cdot XY}$$

And $c_{XY\cdot X} = Q(X,Y,X,X),\ c_{X\cdot XY} = Q(X,X,X,Y)$

$$c_{XY\cdot Y} = Q(X,Y,Y,Y),\ c_{Y\cdot XY} = Q(Y,Y,X,Y)$$

$$\delta_{X\cdot XY} = P(X,X,X,Y),\ \delta_{Y\cdot XY} = P(Y,Y,X,Y)$$

Table 7.2: Under assumptions (A.1) and (A.2), τ_1, τ_2, τ_3 and τ_{XYb} are known functions of σ^2_{XB}, σ^2_{XW}, σ^2_{YB}, σ^2_{YW}, σ_{XYB}, σ_{XYW}, c_{1X}, c_{1XY}, γ_X, γ_{XY}, $c_{X\cdot Y}$, $c_{Y\cdot X}$, $\delta_{X\cdot Y}$, $c_{XY\cdot X}$, $c_{X\cdot XY}$, $\delta_{X\cdot XY}$.

Our approach to estimation is as follows. We estimate the quantities in (i) - (iv) by an ad hoc extension of the usual ANOVA estimation procedure for variance components models (e.g. Searle, 1971). Our estimators are

$$\text{(i)} \quad \hat{\sigma}_{XYW} = \Sigma\, s_{xyi}/n$$

$$\text{where } s_{xyi} = \sum_j (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)/(m_i - 1)$$

$$\bar{x}_i = \sum_j x_{ij}/m_i \;,\; \bar{y}_i = \sum_j y_{ij}/m_i$$

$$\text{(ii)} \quad \hat{\sigma}_{XYB} = (\Sigma\, m_i(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) - (n-1)\,\hat{\sigma}_{XYW})/(m_o - m^*)$$

$$\text{where } \bar{x} = \Sigma\Sigma\, x_{ij}/m_o \;,\; \bar{y} = \Sigma\Sigma\, y_{ij}/m_o$$

$$\text{(iii)} \quad \hat{P}(X,Y,Z,V) = \Sigma\, B_i(X,Y,Z,V)/n - n\,\hat{\sigma}_{XYW}\,\hat{\sigma}_{ZVW}/(n-1)$$

$$\text{where } B_i(X,Y,Z,V) = \alpha_i A_i(X,Y,Z,V) - \beta_i(A_i(X,V,Y,Z) + A_i(X,Z,Y,V))$$

$$A_i(X,Y,Z,V) = s_{xyi}\, s_{zvi}$$

$$\alpha_i = m_i\beta_i + 1/(n-1)$$

$$\beta_i = (m_i-1)/(m_i+1)(m_i-2)$$

$$\text{(iv)} \quad \hat{Q}(X,Y,Z,V) = \left[n\,\Sigma(s_{xyi} - \hat{\sigma}_{XYW})(\bar{z}_i - \bar{z})(\bar{v}_i - \bar{v})/(n-1) - \lambda\hat{P}(X,Y,Z,V)\right]/(n-2)$$

$$\text{where } \lambda = \Sigma\,(m_o - 2m_i)/m_i m_o$$

In Lemma A.2 we show that these estimators are unbiased. We then substitute the various estimates into the formulae in Chapters 5 and 7 to obtain the estimates for Tables 5.2 - 7.2. Since these formulae are in general non-linear the unbiasedness property is generally lost.

In order to prove Lemma A.2 we require the following results.

Lemma A.1

$X_1 \ldots X_4$ are jointly normally distributed random vectors such

that

$$E(X_i) = \mu_i \quad , \quad \text{cov}(X_i, X_j) = \Sigma_{ij} \qquad i,j = 1 \ldots 4$$

Let A and B be conformable matrices (of constants). Then

$$E(X_1'A\, X_2) = \text{tr}(A\, \Sigma_{21}) + \mu_1'A\, \mu_2$$

$$\begin{aligned}\text{cov}(X_1'A\, X_2,\, X_3'B\, X_4) &= \text{tr}(A\, \Sigma_{24}\, B'\Sigma_{31} + A\, \Sigma_{23}\, B\, \Sigma_{41}) \\ &\quad + \mu_1'A\, \Sigma_{24}B'\mu_3 + \mu_2'A'\Sigma_{14}B'\mu_3 \\ &\quad + \mu_1'\, A\, \Sigma_{23}B\, \mu_4 + \mu_2'A'\Sigma_{13}B\mu_4\end{aligned}$$

<u>Proof</u>:

$$\begin{aligned}E(X_1'A\, X_2) &= E\, \text{tr}(X_1'A\, X_2) \\ &= \text{tr}(A\, E\, X_2X_1') \\ &= \text{tr}(A(\Sigma_{21} + \mu_2\mu_1')) \\ &= \text{tr}(A\, \Sigma_{21}) + \mu_1'\, A\, \mu_2\end{aligned}$$

$$\begin{aligned}\text{cov}(X_1'A\, X_2,\, X_3'B\, X_4) &= \text{cov}((X_1 - \mu_1)'A(X_2 - \mu_2),\, (X_3 - \mu_3)'B(X_4 - \mu_4)) \\ &\quad + \text{cov}(\mu_1'A(X_2 - \mu_2),\, \mu_3'B(X_4 - \mu_4)) \\ &\quad + \text{cov}((X_1 - \mu_1)'A\, \mu_2,\, \mu_3'B(X_4 - \mu_4)) \\ &\quad + \text{cov}(\mu_1'A(X_2 - \mu_2),\, (X_3 - \mu_3)'B\, \mu_4) \\ &\quad + \text{cov}((X_1 - \mu_1)'A\, \mu_2\, ,\, (X_3 - \mu_3)'B\, \mu_4) \\ &= \text{cov}((X_1 - \mu_1)'A(X_2 - \mu_2),\, (X_3 - \mu_3)'B(X_4 - \mu_4)) \\ &\quad + \mu_1'A\, \Sigma_{24}B'\, \mu_3 + \mu_2'A'\Sigma_{14}B'\mu_3 \\ &\quad + \mu_1'A\, \Sigma_{23}B\, \mu_4 + \mu_2'\, A'\, \Sigma_{13}\, B\, \mu_4\end{aligned}$$

The result follows by noting that (e.g. Anderson, 1958, p. 39)

$$\begin{aligned}&\text{cov}\left[(X_1 - \mu_1)_i\, (X_2 - \mu_2)_j\, ,\, (X_3 - \mu_3)_k\, (X_4 - \mu_4)_\ell\right] \\ &= \Sigma_{13ik}\, \Sigma_{24j\ell} + \Sigma_{14i\ell}\, \Sigma_{23jk}\end{aligned}$$

Lemma A.2

Under assumptions (A.1) and (A.2) $\hat{\sigma}_{XYW}$, $\hat{\sigma}_{XYB}$, $\hat{P}(X,Y,Z,V)$ and $\hat{Q}(X,Y,Z,V)$ are unbiased for σ_{XYW}, σ_{XYB}, $P(X,Y,Z,V)$ and $Q(X,Y,Z,V)$ respectively.

Proof:

(i) $\hat{\sigma}_{XYW} = \Sigma\ x_i'\ P_{Wm_i}\ y_i/(m_i-1)n$

where $x_i' = (x_{i1} \ldots x_{im_i})$

$y_i' = (y_{i1} \ldots y_{im_i})$

$P_{Wm_i} = I_{m_i} - 1_{m_i} 1_{m_i}'/m_i$

Now $E(x_i|\theta_i) = \mu_{X_i} 1_{m_i}$ (A.3)

$E(y_i|\theta_i) = \mu_{Y_i} 1_{m_i}$ (A.4)

$\text{cov}(x_i, y_i|\theta_i) = \sigma_{XY_i} I_{m_i}$ (A.5)

Hence from Lemma A.1

$$E(\hat{\sigma}_{XYW}|\underline{\theta}) = \Sigma(\text{tr}(\sigma_{XY_i} P_{Wm_i}) + \mu_{X_i} 1_{m_i}' P_{Wm_i} 1_{m_i} \mu_{y_i})/(m_i-1)n$$

$$= \Sigma\ \sigma_{XYi}/n$$

where $\underline{\theta} = (\theta_1 \ldots \theta_n)'$

Hence $E(\hat{\sigma}_{XYW}) = \sigma_{XYW}$

(ii) $\Sigma\ m_i(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) = x'A\ y$

where $x' = (x_1' \ldots x_n')$

$y' = (y_1' \ldots y_n')$

$A = \text{diag}(P_{Bm_i}) - P_{Bm_o}$

$P_{Bm} = I_m - P_{Wm}$

Now $E(x|\underline{\theta}) = (\mu_{X_1} 1_{m_1}' \ldots \mu_{X_n} 1_{m_n}')'$

$E(y|\underline{\theta}) = (\mu_{Y_1} 1_{m_1}' \ldots \mu_{Y_n} 1_{m_n}')$

$\text{cov}(x,y|\underline{\theta}) = \text{diag}(\sigma_{XY_i} I_{m_i})$

Hence from Lemma A.1

$$E(x'Ay|\underline{\theta}) = \text{tr}(\text{diag}(\sigma_{XY_i} P_{Bm_i}) - \text{diag}(\sigma_{XY_i} I_{m_i}) P_{Bm_o})$$

$$+ \Sigma m_i \mu_{X_i} \mu_{Y_i} - (\Sigma\, m_i \mu_{X_i})(\Sigma\, m_i \mu_{Y_i})/m_o$$

$$= \Sigma(1 - m_i/m_o)\sigma_{XY_i} + \underline{\mu}_X' B \underline{\mu}_Y \qquad \text{(A.6)}$$

where $\underline{\mu}_X' = (\mu_{X_1} \ldots \mu_{X_n})$

$\underline{\mu}_Y' = (\mu_{Y_1} \ldots \mu_{Y_n})$

$B_{ij} = m_i - m_i^2/m_o \qquad \text{if } i=j$

$= -m_i m_j/m_o \qquad \text{if } i \neq j$

Now $E(\underline{\mu}_X) = \mu_X 1_n$

$E(\underline{\mu}_Y) = \mu_Y 1_n$

$\text{cov}(\underline{\mu}_X, \underline{\mu}_Y) = \sigma_{XYB} I_n$

Hence from Lemma A.1 and (A.6)

$$E(x'A_B\, y) = \Sigma(1 - m_i/m_o)\sigma_{XYW} + \text{tr}(B\, \sigma_{XYB} I_n)$$

$$+ \mu_X \mu_Y 1_n' B\, 1_n$$

$$= (n-1)\sigma_{XYW} + (m_o - m^*)\sigma_{XYB} \qquad \text{(A.7)}$$

Substituting (A.7) into $E(\hat{\sigma}_{XYB})$ we obtain

$$E(\hat{\sigma}_{XYB}) = \sigma_{XYB}$$

(iii) From (i)

$$s_{xy_i} = x_i' P_{Wm_i}\, y_i/(m_i - 1)$$

Hence from (A.3) - (A.5) and Lemma A.1 we have

$$\mathrm{cov}(s_{xy_i}, s_{zv_i} | \theta_i) = \mathrm{tr}(P_{Wm_i}(\sigma_{XV_i}\sigma_{YZ_i} + \sigma_{XZ_i}\sigma_{YV_i})/(m_i-1)^2)$$

since $P_{Wm_i} 1_{m_i} = 0$

$$= (\sigma_{XV_i}\sigma_{YZ_i} + \sigma_{XZ_i}\sigma_{YV_i})/(m_i-1)$$

Hence

$$E(s_{xy_i}, s_{zv_i}) = \mathrm{cov}(s_{xy_i}, s_{zv_i}) + \sigma_{XYW}\sigma_{ZVW}$$

$$= E(\mathrm{cov}(s_{xy_i}, s_{zv_i} | \theta_i)) + \mathrm{cov}(\sigma_{XY_i}, \sigma_{ZV_i}) + \sigma_{XYW}\sigma_{ZVW}$$

$$= (P(X,V,Y,Z) + P(X,Z,Y,V) + \sigma_{XYW}\sigma_{YZW} + \sigma_{XZW}\sigma_{YVW})/(m_i-1)$$

Hence

$$E(B_i(X,Y,Z,V)) = E(A_i(X,Y,Z,V)/(n-1)) + \beta_i(m_i-2)/(m_i-1)(P(X,Y,Z,V) + \sigma_{XYW}\sigma_{ZVW}) + \beta_i(m_i/(m_i-1) - 1 - 1/(m_i-1)).$$

$$(P(X,V,Y,Z) + P(X,Z,Y,V) + \sigma_{XYW}\sigma_{YZW} + \sigma_{XZW}\sigma_{YVW})$$

$$= E(A_i(X,Y,Z,V)/(n-1)) + P(X,Y,Z,V) + \sigma_{XYW}\sigma_{ZVW}$$

Hence

$$E(\hat{P}(X,Y,Z,V)) = P(X,Y,Z,V) + \sigma_{XYW}\sigma_{ZVW} + E(\Sigma A_i(X,Y,Z,V) - \Sigma\Sigma s_{xy_i} s_{zv_j})/n(n-1)$$

$$= P(X,Y,Z,V) + \sigma_{XYW}\sigma_{ZVW} + E(\Sigma\Sigma_{i\neq j} s_{xy_i} s_{zv_j})/n(n-1)$$

$$= P(X,Y,Z,V)$$

(iv) Within clusters the cluster means $\bar{z}_i$ and $\bar{v}_i$ are independent of the cluster covariances s_{xy_i}. Hence

$$E(\hat{Q}(X,Y,Z,V)|\underline{\theta}) = [n\ \Sigma(\sigma_{XY_i} - \Sigma\ \sigma_{XY_j}/n)E((\bar{z}_i - \bar{z})(\bar{v}_i - \bar{v})|\underline{\theta}) /(n-1) - \lambda\ E(\hat{P}(X,Y,Z,V)|\underline{\theta})]/(n-2) \qquad (A.8)$$

Now

$$(\bar{z}_i - \bar{z})(\bar{v}_i - v) = z'A^{(i)}V$$

where
$$A^{(i)}_{jk,j'k'} = (m_o - m_i)^2/m_o^2 m_i^2 \quad \text{if } j=i \quad j'=i$$
$$= -(m_o - m_i)/m_o^2 m_i \quad \text{if } j=i \quad j'\neq i$$
$$\text{or} \quad \text{if } j\neq i \quad j'=i$$
$$= -1/m_o^2 \quad \text{if } j\neq i \quad j'\neq i$$

Hence from Lemma A.1

$$E((\bar{z}_i - \bar{z})(\bar{v}_i - \bar{v})|\underline{\theta}) = tr(A^{(i)}\ diag(\sigma_{zv_i}\ I_{m_i}))$$
$$+ \mu_{z_i}\mu_{v_i}(m_o - m_i)^2/m_o^2$$
$$- \mu_{z_i} \sum_{j\neq i} \mu_{v_j} m_j(m_o - m_i)/m_o^2$$
$$- \mu_{v_i} \sum_{j\neq i} \mu_{z_j} m_j(m_o - m_i)/m_o^2$$
$$+ \sum_{j,k\neq i} \mu_{z_j}\mu_{v_k} m_j m_k/m_o^2$$
$$= \sigma_{zv_i}(m_o - m_i)^2/m_i m_o^2 + \sum_{j\neq i} m_j\sigma_{zv_j}/m_o^2$$
$$+ \mu_{z_i}\mu_{v_i}(m_o^2 - 2\ m_i m_o)/m_o^2$$
$$+ (\Sigma\ \mu_{z_j} m_j)(\Sigma\ \mu_{v_j} m_j)/m_o^2$$
$$+ \text{cross-product terms in } \mu_{X_i} \text{ and } \mu_{Y_j} (i\neq j)$$

(A.9)

Now, when (A.9) is substituted into (A.8), the term $\Sigma \mu_z m \ \ \Sigma \mu_v m/m_o^2$ disappears since it does not depend on i. Also the cross-product terms disappear when the expectation of (A.8) is taken (note that without loss of generality we may assume $\mu_Z = \mu_V = 0$ since $\bar{z}_i - \bar{z}$ and $\bar{v}_i - \bar{v}$ do not depend on μ_Z or μ_V). Hence

$$E(\hat{Q}(X,Y,Z,V)) = \Big[n \sum_i E(\sigma_{XYi} - \sum_j \sigma_{XY_j}/n)(\sigma_{ZV_i}(m_o-m_i)^2/m_i m_o + \sum_{j=i} m_j\sigma_{ZV_j}/m_o^2 + \mu_{Z_i}\mu_{V_i}(m_o - 2m_i)/m_o)/(n-1) - \lambda P(X,Y,Z,V)\Big]/(n-2)$$

$$= \Big[n \sum_i E(\sigma_{XY_i} - \sum_j \sigma_{XY_j}/n)(\sigma_{ZV_i}(m_o - 2m_i)/m_i m_o + \mu_{Z_i}\mu_{V_i}(m_o - 2m_i)/m_o)/(n-1) - \lambda P(X,Y,Z,V)\Big]/(n-2)$$

$$= \Big[\Sigma(m_o - 2m_i)P(X,Y,Z,V)/m_i m_o + \Sigma(m_o - 2m_i)Q(X,Y,Z,V)/m_o - \lambda\, P(X,Y,Z,V)\Big]/(n-2)$$

$$= Q(X,Y,Z,V)$$

Finally we consider the estimation of misspecification effects. We consider three broad procedures.

(i) Making assumptions (A.1) and (A.2), we may substitute the above estimates into the various formulae.

(ii) Making the weaker assumption B we may alternatively estimate the misspecification effect of T_{Ym} using the following result.

Lemma A.3

Let $\hat{\sigma}_X^2 = \Big[\Sigma m_i(\bar{x}_i-\bar{x})^2 + \Sigma(m_i-1)(1-m_i/m_o)s_{X_i}^2\Big]/(m_o-m^*)$

Then under Assumption B $E_I(\hat{\sigma}_X^2 | s,\underline{M}) = \sigma_X^2$

Proof

From Lemma A.2(ii)

$E_I(\Sigma m_i(\bar{x}_i - \bar{x})^2 | \underline{\theta}) = \Sigma(1-m_i/m_o)\sigma_{Xi}^2 + \underline{\mu}_X' B \underline{\mu}_X$

Hence under B

$$E_I(\Sigma m_i(\bar{x}_i-\bar{x})^2 \mid s,\underline{M}) = \Sigma(1-m_i/m_o)\ \sigma_{Xw}^2(M_i) + \Sigma m_i(1-m_i/m_o)\sigma_B^2(M_i)$$

Hence

$$E_I(\hat{\sigma}_X^2 \mid s,\underline{M}) = \left[\Sigma(1-m_i/m_o)\sigma_{Xw}^2(M_i) + \Sigma m_i(1-m_i/m_o)\sigma_B^2(M_i) + \Sigma(m_i-1)(1-m_i/m_o)\ \sigma_{Xw}^2(M_i)\right]/(m_o-m^*)$$

$$= \Sigma m_i(1-m_i/m_o)\ \sigma_X^2/(m_o-m^*)$$

since $\sigma_X^2 = \sigma_{Xw}^2(M_i) + \sigma_{XB}^2(M_i)$

$$= \sigma_X^2$$

An estimate of the misspecification effect of T_{Xm} is then, from Lemma 5.12.

$$\hat{meff}\ (T_{Xm} \mid s,\underline{M},\psi) = 1 + \Sigma m_i(m_i-1)(\hat{\sigma}_X^2-s_{xi}^2)/m_o\hat{\sigma}_X^2$$

Estimates of the misspecification effects of T_{Xv} and T_{XYc} may then be obtained by replacing X_{ij} by $(X_{ij}-\hat{\mu}_X)^2$ and $(X_{ij}-\hat{\mu}_X)(Y_{ij}-\hat{\mu}_Y)$ respectively in the above formula. The justification for this approach derives from comparing equations (5.42) and (5.91) with equation (5.28). Estimates of the misspecification effects of T_{XYr} and T_{XYb} may be obtained similarly by using Lemma 5.9. This does require the evaluation of some partial derivatives.

(iii) We may avoid the evaluation of partial derivatives in (ii) by using Lemma 5.10. The misspecification effects are given in (7.3) and (7.23) in terms of r_i, defined in (7.8), and b_i, defined in (7.27). A possible estimate of $(\mu_{Xi}-\mu_X)(\mu_{Yi}-\mu_Y) + \sigma_{XYi}$, the numerator of r_i, is

$$\sum_{j=1}^{m_i} (x_{ij}-\hat{\mu}_X)(Y_{ij}-\hat{\mu}_Y)/m_i$$

Estimates of $(\mu_{Xi}-\mu_X)^2 + \sigma_{Xi}^2$ and $(\mu_{Yi}-\mu_Y)^2 + \sigma_{Yi}^2$ may be defined similarly.

A comparison of estimates obtained by methods (i) and (ii) above for the misspecification effects of T_{Xv} and T_{XYc} are given in Table A.1.

Table A.1 : Estimates for FES data

	$\hat{meff}(T_{Xv})$			$\hat{meff}(T_{XYc})$		
Variables*	1	2	3	1,2	1,3	2,3
Method (i)	1.137	2.112	1.300	1.615	1.303	1.660
Method (ii)	1.174	1.425	1.284	1.420	1.286	1.452

*variable 1 = log(V1) etc.

REFERENCES

Abramowitz, M. and Stegun, I.A. (1964) Handbook of Mathematical Functions. Dover, New York.

Adhikari, B.P. and Sarma, Y.R. (1978) Some tests of hypotheses for cluster sampling. Sankhya B 40 29-37.

Ahmavaara, Y. (1954) The mathematical theory of factorial invariance under selection. Psychometrika 19 27-38.

Aitkin, A.C. (1934) Note on selection from a multivariate normal population. Proc. Edin. Math. Soc. 4 106-110.

Aitkin, A.C. (1935) A further note on multivariate selection. Proc. Edin. Math. Soc. 5 37.

Altham, P.M.E. (1976) Discrete variable analysis for individuals grouped into families. Biometrika 63 263-269.

Anderson, T.W. (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J. Amer. Statist. Assoc. 52 200-203.

Anderson, T.W. (1958) An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Anderson, T.W. (1963) Asymptotic theory for principal components analysis. Ann. Math. Statist. 34 122-48.

Aoyama, H. (1954) A study of stratified random sampling. Ann. Inst. Statist. Math. 6 1-36.

Arnold, S.F. (1981) The Theory of Linear Models and Multivariate Analysis. Wiley, New York.

Barnard, G.A. (1971) Discussion of a paper by V.P. Godambe and M.E. Thompson. Bayes, Fiducial and Frequentist Aspects of Regression Analysis in Survey Sampling. J.Roy. Statist. Soc. B 33 361-390.

Barndorff-Nielsen, O. (1978) Information and Exponential Families. Wiley, Chichester.

Bartholemew, D.J. (1973) Stochastic Models for Social Processes. 2nd Ed. Wiley, London.

Bebbington A.C. and Smith T.M.F. (1977) The effect of survey design on multivariate analysis. In O'Muircheartaigh and Payne (1977b).

Bellhouse, D.R., Thompson, M.E. and Godambe, V.P. (1977) Two-stage sampling with exchangeable prior distributions. Biometrika 64 97-103.

Bielby, W.T. (1981) Neighbourhood effects : a LISREL model for clustered samples. Sociological Methods and Research 10, 82-111.

Birnbaum,Z.W., Paulson, E. and Andrews, F.C. (1950) On the effect of selection performed on some co-ordinates of a multidimensional population. Psychometrika 15, 191-204.

Blalock, H.M. (1968) Theory building and causal inference. In 'Methodology in Social Research' Eds. H.M. Blalock and A.B. Blalock. McGraw-Hill, New York.

Bloxom, B. (1972) Alternative approaches to factorial invariance. Psychometrika 29 187-206.

Booth, G. and Sedransk,J. (1969) Planning some two factor comparative surveys. J.Amer.Statist. Assoc. 64 560-573.

Brewer, K.R.W. (1963) Ratio estimation and finite populations : some results deducible from the assumption of an underlying stochastic process. Aust. J. Statist. 5 93-105.

Brewer, K.R.W., Foreman, E.K., Mellor, R.W. and Trewin, D.J. (1977) Use of experimental design and population modelling in survey sampling. Bull. Int. Statist. Inst. 47 Book 3. 173-190.

Brewer, K.R.W. and Mellor, R.W. (1973) The effect of sample structure on analytical surveys. Aust. J. Statist. 15 145-152.

Brier, S.S. (1980) Analysis of contingency tables under cluster sampling. Biometrika 67 591-596.

Brown, K.G. (1978) Estimation of variance components using residuals J. Amer. Statist. Assoc. 73 141-146.

Campbell, C. (1977) Properties of ordinary and weighted least squares estimators of regression coefficients for two-stage samples. Amer. Statist. Assoc: Proc. of the Social Statistics Section. 800-805.

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1976) Some results on generalised difference estimation and generalised regression estimation for finite populations. Biometrika 63 615-620.

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. Wiley, New York.

Chaudhuri, A. (1978) On estimating the variance of a finite population Metrika 25, 65-76.

Chen, C.F. (1979) Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. J.Roy.Statist. Soc. B 41,235-248.

Cochran, W.G. (1946) Relative accuracy of systematic and stratified random samples for a certain class of population. Ann. Math. Statist. 17 164-177.

Cochran, W.G. (1963) Sampling Techniques 2nd Ed. Wiley, New York.

Cochran, W.G. (1977) Sampling Techniques 3rd Ed. Wiley, New York.

Cohen, J.E. (1976) The distribution of the chi-squared statistic under cluster sampling from contingency tables. J.Amer. Statist. Assoc. 71 665-670.

Coleman, J.S. (1959) Relational analysis : a study of social organisation with survey methods. Human Organisation 17 28-36.

Cowan, J. and Binder, D. (1978) The effect of a two-stage sample design on tests of independence of a 2x2 table. Survey Methodology 4 29-56.

Cox, D.R. and Hinkley, D.V. (1974) Theoretical Statistics. Chapman and Hall, London.

Das, A.K. and Tripathi, T.P. (1977) Admissible estimators for quadratic forms in finite populations. Bull.Int.Statist. Inst. 47 Book 4 132-135.

Das, A.K. and Tripathi, T.P. (1978) Use of auxiliary information in estimating the finite population variance. Sankhyā C 40 139-148.

Davis, A.W. (1977) Asymptotic theory for principal components analysis: non-normal case. Aust. J. Statist. 19 206-212.

Demets, D. and Halperin, M. (1977) Estimation of a simple regression coefficient in samples arising from a subsampling procedure. Biometrics 33 47-56.

Dempster, A.P. (1969) Elements of Continuous Multivariate Analysis. Addison-Wesley. Reading, Mass.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. J.Roy.Statist. Soc. B 39 1-38.

Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981) Estimation in covariance components models. J.Amer.Statist. Assoc. 76 341-353.

Des Raj (1958) On the relative accuracy of some sampling techniques. J. Amer. Statist. Assoc. 53 98-101.

Dogan, M. and Rokkam, S. (Eds) (1969) Quantitative Ecological Analysis in the Social Sciences. MIT Press, Cambridge, Mass.

Fellegi, I.P. (1980) Approximate tests of independence and goodness of fit based on stratified multistage samples. J.Amer.Statist.Assoc. 75 261-268.

Ferber, R. (1980) Readings in the analysis of survey data. American Marketing Assoc., Chicago.

Fielding, A. (1977) Latent structure models. In O'Muircheartaigh and Payne (1977a) 125-158.

Fields, J.M. (1971) The sample cluster : a neglected data source. Public Opinion Quarterly 34, 593-603.

Foreman, E.K. and Brewer, K.R.W. (1971) The efficient use of supplementary information in standard sampling procedures J.Roy.Statist.Soc. B 33 391-400.

Frankel, M.R. (1971) Inference from Survey Samples. Institute for Social Research, Ann Arbor.

Freeman, D.H. and Brock, D.B. (1978) The role of covariance matrix estimation in the analysis of complex survey data. In Namboodiri (1978).

Freeman, D.H., Freeman, J.L. and Brock, D.B. (1977) Modularization for the analysis of complex sample survey data. Bull.Int.Statist.Inst. Book 3, 3-20.

Freeman, D.H., Freeman, J.L. Brock, D.B. and Koch, G.G. (1976) Strategies in the multivariate analysis of data from complex surveys II. Int. Statist.Rev. 44 317-330.

Freeman, D.H. and Koch, G.G. (1976) An asymptotic covariance structure for testing hypotheses on raked contingency tables from complex sample surveys. Amer.Stat.Assoc. Proc. of Social Statistics Sect. 330-335.

Fujikoshi, Y. (1980) Asymptotic expansions for the distributions of the sample roots under non-normality. Biometrika 67 45-51.

Fuller, W.A. (1973) Regression analysis for sample surveys. Int. Assoc. of Survey Statisticians. 1st Meeting, Vienna, August, 1973.

Fuller, W.A. (1975) Regression analysis for sample surveys. Sankhyá C 37 117-132.

Fuller, W.A. (1982) Regression estimation of regression equations. Manuscript.

Fuller, W.A. and Battese, G.E. (1973) Transformations for estimation of linear models with nested error structure. J.Amer.Statist. Assoc. 68 626-632.

Fuller, W.A., Pantula, S.G. and Amemiya, Y. (1982) The covariance matrix of estimators for the factor model. Manuscript.

Galtung, J. (1967) Theory and Methods of Social Research. Allen and Unwin, London.

Girschick, M.A. (1939) On the sampling theory of roots of deterinmental equations. Ann.Math.Statist. 10 203-224.

Gnanadesikan, R. (1977) Methods for Statistical Data Analysis of Multivariate Observations. Wiley, New York.

Goldberger, A.S. (1981) Linear regression after selection. J.Econometrics 15 357-366.

Gosnell, H.F. and Schmidt, M.J. (1936) Factorial and Correlational analysis of the 1934 vote in Chicago. J.Amer.Statist.Assoc. 31 507-518.

Graybill, W.A. (1954) On quadratic estimates of variance components Ann. Math. Statist. 25 367-372.

Grimes, C.E. and Sukhatme, B.V. (1980) A regression-type estimator based on preliminary test of significance. J.Amer.Statist. Assoc. 75 957-962.

Gupta, J.P., Singh, R. and Lal, B. (1978) On the estimation of the finite population correlation coefficient I. Sankhya C 40 38-59.

Gupta, J.P., Singh, R. and Lal, B. (1979) On the estimation of the finite population correlation coefficient II. Sankhya C 41. 1-39.

Hájek, J. (1971) Contribution to discussion of paper by D. Basu. In Foundations of Statistical Inference Eds. V.P. Godambe and D.A. Sprott p.236. Holt, Rinehartand Winston, Toronto.

Hall, P. and Heyde, C.C. (1980) Martingale limit theory and its application. Academic Press, New York.

Hansen, M.H. and Hurwitz, W.N. (1943) On the theory of sampling from finite populations. Ann.Math.Statist. 14 333-362.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953) Sample Survey Methods and Theory Vol.I. Wiley, New York.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1978) Foundations of inference in survey sampling. Amer.Statist.Assoc. Proc. of Sect. on Survey Research Methods. 82-107.

Hartley, H.O. and Rao, J.N.K. (1967) Maximum likelihood estimation for the mixed analysis of variance model. Biometrika 54 93-108.

Hartley, H.O. and Sielken, R.L. (1975) A 'superpopulation viewpoint' for finite population sampling. Biometrics 31 411-422.

Harville, D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. J.Amer.Statist.Assoc. 72 320-340.

Heckman, J.J. (1979) Sample selection bias as a specification error. Econometrics 47 153-161.

Hemmerle, W.J. and Hartley H.O. (1973) Computing ML estimates for the mixed A.O.V. Model using the W transformation. Technometrics 15 819-831.

Hendricks, W.A. (1944) The relative efficiencies of groups of farms as sampling units. J.Amer.Statist.Assoc. 39 366-376.

Hendry, D.F. and Mizon, G.E. (1978) Serial correlation as a convenient simplification not a nuisance. Economic J. 88 549-563.

Hess, J.L. (1979) Sensitivity of MINQUE with respect to a priori weights. Biometrics 35 645-649.

Hill, B.M. (1980) Robust analysis of the random model and weighted least squares regression. In 'Evaluation of Econometric Models' J. Kmenta and J.B. Ramsey (Eds) Academic Press, New York.

Hinkley, D.V. (1978) Predictive likelihood. Ann.Statist. 7 718-728.

Holt, D. (1980) Discussion of a paper by Verma, V.J., Scott C. and O'Muircheartaigh, C.A. Sample designs and sampling errors for the World Fertility Survey. J.Roy.Statist.Soc.A. 143 431-463.

Holt, D., Richardson, S.C. and Mitchell, P.W. (1976) The analysis of correlations in complex survey data. Manuscript.

Holt, D. and Scott, A.J. (1981) Regression analysis using survey data. The Statistician 30 169-178.

Holt, D., Scott, A.J. and Ewings, P.D. (1980a) Chi-squared tests with survey data. J.Roy.Statist.Soc.A 143 302-320.

Holt, D. and Smith, T.M.F. (1976) The design of surveys for planning purposes. Aust.J.Statist. 18 37-44.

Holt, D. Smith, T.M.F. and Winter, P.D. (1980b) Regression analysis of data from complex surveys. J.Roy.Statist.Soc.A 143 474-487.

Horvitz, D.G. and Thompson, D.J. (1952) A generalisation of sampling without replacement from a finite universe. J.Amer.Statist. Assoc. 47 663-685.

Huber, P.J. (1972) Robust statistics : a review. Ann.Math.Statist. 43 1041-1067.

Imrey, P.B., Francis, M.E. and Sobel, E. (1979) Analysis of categorical data obtained by stratified random sampling I. Commun.Statist A8 653-670.

Imrey, P.B., Francis, M.E. and Sobel, E. (1980) Modelling contingency tables from complex surveys. Amer.Statist.Assoc. Proc. of Survey Methods Sect. 212-217.

Isaki, C.T. and Fuller, W.A. (1982) Survey design under the regression superpopulation model. J.Amer.Statist.Assoc. 77 89-96.

James, A.T. (1960) The distribution of the roots of the covariance matrix. Ann.Math.Statist. 31 151-158.

James, A.T. (1964) Distributions of matrix variates and latent roots of the covariance matrix. Ann. Math. Statist. 35 475-501.

Jessen, R.J. (1942) Statistical investigation of a sample survey for obtaining farm facts. Iowa Agricultural Station Research Bulletin No. 304.

Johnson, N.L. and Kotz, S. (1972) Distributions in Statistics: Continuous Multivariate Distributions. Wiley, New York.

Johnson, N.L. and Smith, H. (Eds) (1969) New Developments in Survey Sampling. Wiley, New York.

Jonrup, H. and Rennermalm, B. (1976) Regression analysis in samples from finite populations. Scand.J.Statist. 3 33-37.

Joreskog, K.G. and Goldberger, A.S. (1975) Estimation of a model with multiple indicators and multiple causes of a single latent variable. J. Amer.Statist.Assoc. 70 631-639.

Joreskog, K.G. and Sorbom, D. (1978) LISREL IV Users Guide. International Educational Services, Chicago.

Kalton, G. (1976) Discussion of Smith (1976).

Kalton, G. (1977) Practical methods for estimating survey sampling errors. Bull.Int.Statist.Inst. 47 Book 3 495-512.

Kempthorne, O. (1978) Discussion of Hansen et al (1978).

Kemsley, W.F.F. (1969) Family Expenditure Survey: Handbook on the sample, fieldwork and coding procedures. HMSO, London.

Kendall, M.G. and Stuart, A. (1969) The Advanced Theory of Statistics Vol.I 3rd Ed. Hafner, New York.

Kish, L. (1965) Survey Sampling. Wiley, New York.

Kish, L. (1969) Design and estimation for subclasses, comparisons and analytical statistics. In Johnson and Smith (1969).

Kish, L. and Frankel, M.R. (1970) Balanced repeated replication for standard errors. J.Amer.Statist.Assoc. 65 1071-1094.

Kish, L. and Frankel, M.R. (1974) Inference from complex samples. J.Roy.Statist.Soc.B 36 1-37.

Kish, L. Groves, R.M., and Krotki, K.P. (1976) Sampling Errors for Fertility Surveys. World Fertility Survey Occasional Paper No. 17.

Koch, G.G., Freeman, D.H. and Freeman, J.L. (1975) Strategies in the multivariate analysis of data from complex surveys. Int. Statist. Rev. 43 59-78.

Koch, G.G. and Lemershow, S. (1972) An application of multivariate analysis to complex sample survey data. J.Amer.Statist.Assoc. 67, 780-782.

Konijn, H. (1962) Regression analysis in sample surveys. J.Amer. Statist. Assoc. 57 590-605.

Koop, J.C. (1970) Estimation of correlation for a finite universe. Metrika 15, 105-109.

Koopmans, T.C. (1947) Measurement without theory. Rev.Econ.Statist. 29 161-172.

Krewski, D. and Rao, J.N.K. (1981) Inference from stratified samples: properties of the linearisation, jacknife and balanced repeated replication methods. Ann.Statist. 9 1010-1019.

Krishnaiah, P.R. and Chattopadhyay, A.K. (1975) On some non-central distributions in multivariate analysis. S.Afr.Statist.J. 9 37-46.

Krishnaiah, P.R. and Lee, J.C. (1979) On the asymptotic joint distribution of certain functions of the eigenvalues of four random matrices J. Mult.Anal. 9 248-258.

Krzanowski, N.Z. (1979) Between-groups comparison of principal components. J.Amer.Statist.Assoc. 74 703-707.

Laird, N. (1978) Nonparametric maximum likelihood estimation of a mixing distribution. J.Amer.Statist.Assoc. 73 805-811.

Lamotte, L.R. (1973) Quadratic estimation of variance components. Biometrics 29 311-330.

Lauritzen, S.L. (1974) Sufficiency, prediction and extreme models. Scand.J.Statist. 1 128-134.

Lawley, D.N. (1943) A note on Karl Pearson's Selection formula. Proc. Roy. Soc. Edin.Sect.A. 62 28-30.

Lawley, D.N. (1956) Tests of significance for the latent roots of the covariance and correlation matrix. Biometrika 43 123-141.

Lawley, D.N. and Maxwell, A.E. (1971) Factor analysis as a statistical method. Butterworths, London.

Leamer, E.E. (1978) Specification Searches. Wiley, New York.

Ledermann, W. (1938a) Sampling distribution and selection in a normal population. Biometrika 30 295-304.

Ledermann, W. (1938b) Note on Professor Godfrey H. Thom son's article "The influence of univariate selection on factorial analysis of ability" Brit. J.Psychol. 29 69-73.

Lemeshow, S. (1977) Estimating the variance of the slope of a linear regression in a stratified random sample with the balanced half-sample technique. Amer.Statist.Assoc.Proc. of Social Statist. Sect. 831-836.

Lepkowski, J.M. and Landis, J.R. (1980) Design effects for linear contrasts of proportions and logits. Amer.Statist.Assoc.Proc. of Survey Methods Sect. 224-229.

Liao and Sedransk, J. (1975) Sequential sampling for the comparison of domain means. Biometrika 62 690-693.

Lindley, D.V. and Smith, A.F.M. (1972) Bayes estimates for the linear model. J.Roy.Statist.Soc. B 34 1-41.

Liu, T.P. (1974a) Bayes estimation for the variance of a finite population. Metrika 21 127-132.

Liu, T.P. (1974b) A general unbiased estimator for the variance of a finite population. Sankhya C 36 23-32.

Lord, F.M. and Novick, M.R. (1968) Statistical Theories of Mental Test Scores. Addison-Wesley; Reading, Mass.

Mahalanobis, P.C. (1944) On large-scale sample surveys. Phil. Trans. 231B 329-451.

Mallows, C.L. (1961) Latent vectors of random symmetric matrices. Biometrika 48 133-149.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) Multivariate Analysis. Academic Press, London.

Mathai, A.M. (1980) Moments of the trace of a non-central Withart matrix. Comm. Statist. A9 795-801.

Meredith, W. (1964a) Notes on factorial invariance. Psychometrika 29 177-185.

Meredith, W. (1964b) Rotation to achieve factorial invariance. Psychometrika 29 187-206.

Mkai, C.P.B. (1981) Regression analysis of Family Expenditure Survey data. M.Sc. dissertation, University of Southampton.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) Introduction to the Theory of Statistics. 3rd Ed. McGraw-Hill, New York.

Morgan, J.N. and Sonquist, J.A. (1963) Problems in the analysis of survey data, and a proposal. J.Amer.Statist.Assoc. 58 415-434.

Morrison, D.F. (1971) Expectations and variances of maximum likelihood estimates of the multivariate normal distribution parameters with missing data. J.Amer.Statist.Assoc. 66 602-604.

Morrison, D.F. (1976) Multivariate Statistical Methods 2nd Ed. McGraw Hill, New York.

Muirhead, R.J. (1978) Latent roots and matrix variates : A review of some asymptotic results. Ann.Statist. 6 5-33.

Mukhopadhyay, P. (1978) Estimating the variance of a finite population under a superpopulation maodel. Metrika 25 115-122.

Murthy, M.N. (1963) Generalised unbiased estimation in sampling from finite populations. Sankhya B 25 245-262.

Muthén, B. (1978) Contributions to factor analysis of dichotomous variables. Psychometrika 43 551-560.

Muthén, B. (1981) Factor analysis of dichotomous variables: American attitudes towards abortion. In 'Factor Analysis and Measurements in Sociological Research' Eds. D.J. Jackson and E.F. Borgatta. Sage, London. 201-214.

Namboodiri, N.K. (1978) Survey Sampling and Measurement. Academic Press, New York.

Nathan, G. (1969) Tests of independence in contingency tables from stratified samples. In Johnson and Smith (1969).

Nathan, G. (1972) On asymptotic power of tests for independence in contingency tables from complex stratified samples. J.Amer. Statist. Assoc. 67 917-920.

Nathan, G. (1975) Tests of independence in contingency tables from stratified proportional samples. Sankhya C 37 77-87.

Nathan, G. (1982) A simulation comparison of estimators for a regression coefficient under differential non-response. Paper presented at International Meeting on Analysis of Sample Survey Data, Jerusalem, June 1982.

Nathan, G. and Holt, D. (1980) The effect of survey designs on regression analysis. J.Roy. Statist. Soc.B. 42 377-386.

Novick, M.R. and Jackson, P.H. (1974) Statistical Methods for Educational and Psychological Research. McGraw Hill, New York.

O'Muircheartaigh, C.A. (1977) Statistical analysis in the context of survey research. In O'Muircheataigh and Payne (1977a)

O'Muircheartaigh, C.A. and Payne, C. (1977a) The Analysis of Survey Data Vol. I: Exploring Data Structures. Wiley, New York.

O'Muircheartaigh, C.A. and Payne, C. (1977b) The Analysis of Survey Data Vol. 2: Model Fitting. Wiley, New York.

Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58 545-554.

Pearson, K. (1903) On the influence of natural selection on the variability and correlation of organs. Phil.Trans.A. 200 1-66.

Pearson, K. (1912) On the general theory of the influence of selection on correlation and variation. Biometrika 8 437-443.

Pfefferman, D. and Nathan, G. (1981) Regression analysis of data from a cluster sample. J.Amer.Statist. Assoc. 76 681-689.

Porter, R.M. (1973) On the use of sample survey weights in the linear model. Ann.Econ.Social Meas. 2 141-158.

Praetz, P. (1981) A note on the effect of autocorrelation on multiple regression statistics. Aust. J. Statist. 23 309-313.

Proctor, C.H. (1980) Estimating Smith's b from sample survey data. Amer.Statist. Assoc. Proc. Survey Res. Methods Sect. 761-765.

Pukelsheim, F. (1976) Estimating variance components in linear models. J. Mult. Anal. 6 626-629.

Pukelsheim, F. (1977) On Hsu's model in regression analysis. Math. Operations forsch. Statist. Ser.Statistics 8 323-331.

Rao, C.R. (1971a) Estimation of variance and covariance components - MINQUE theory. J.Mult.Anal. 1 257-275.

Rao, C.R. (1971b) Minimum variance quadratic unbiased estimation of variance components. J.Mult.Anal. 1 445-456.

Rao, C.R. (1973) Linear Statistical Inference and its Applications. 2nd Ed. Wiley, New York.

Rao, C.R. (1979) MINQE theory and its relation to ML and MML estimation of variance components. Sankhya B 41. 138-153.

Rao, J.N.K. (1973) On double sampling for stratification in analytical surveys. Biometrika 60 125-133.

Rao, J.N.K. (1975a) Analytic studies of sample survey data. Survey Methodology 1. Supplementary Issue 1-76.

Rao, J.N.K. (1975b) Sampling designs involving unequal probabilities of selection. Appendix 1. Invited paper at Int.Assoc. of Survey Statisticians Meeting, Warsaw.

Rao, J.N.K. and Scott, A.J. (1981) The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. J.Amer. Statist. Assoc. 76 221-230.

Rao, P.S.R.S., Kaplan, J. and Cochran, W.G. (1981) Estimators for the one-way random effects model with unequal error variances. J.Amer.Statist.Assoc. 76 89-97.

Rao, T.J. (1967) On the cjoice of a strategy for the ratio methods of estimation. J.Roy.Statist. Soc.B. 29 392-397.

Royall, R.M. (1970) On finite population sampling theory under certain linear regression models. Biometrika 57 377-389.

Royall, R.M. (1971) Linear regression models in finite population sampling theory. In V.P. Godambe and D.A. Sprott (Eds) Foundations of Statistical Inference. Holt, Rinehart and Winston, Toronto. 259-274.

Royall, R.M. (1976a) Likelihood functions in finite population sampling theory. Biometrika 63 605-614.

Royall, R.M. (1976b) The linear least-squares prediction approach to two-stage sampling. J.Amer.Statist.Assoc. 71 657-664.

Royall, R.M. and Cumberland, W.G. (1981) An empirical study of the ratio estimator and estimators of its variance. J.Amer.Statist. Assoc. 76 66-88.

Royall, R.M. and Herson, J. (1973a) Robust estimation in finite populations I. J.Amer. Statist. Assoc. 68 880-889.

Royall, R.M. and Herson, J. (1973b) Robust estimation in finite populations II. J.Amer. Statist. Assoc. 68 890-893.

Rubin, D.B. (1977) Formalising subjective notions about the effect of non-respondents in sample surveys. J.Amer.Statist. Assoc. 72 538-543.

Rutter, M. Maughan, B, Mortimore, P. and Ouston, J. (1979) Fifteen Thousand Hours. Open Books, Somerset.

Särndal, C.E. (1980) On Π-inverse weighting versus best linear unbiased weighting in probability sampling. Biometrika 67 639-650.

Scott, A.J. (1977) On the problem of randomisation in survey sampling. Sankhya C 39 1-9.

Scott, A.J. and Smith T.M.F. (1969) Estimation in multi-stage surveys. J.Amer. Statist. Assoc. 68 880-889.

Searle, S.R. (1956) Matrix methods in variance and covariance components analysis. Ann.Math.Statist. 27 737-748.

Searle, S.R. (1971) Linear Models. Wiley, New York.

Searle, S.R. (1979) Maximum likelihood and minimum variance estimation of variance components. In Van Vleck and Searle (1979).

Sedransk, J. (1965a) A double sampling scheme for analytical surveys. J. Amer. Statist. Assoc. 60 985-1004.

Sedransk, J. (1965b) Analytical surveys with cluster sampling. J.Roy. Statist. Soc. B 27 264-278.

Sedransk J. (1967) Designing some multifactor analytical studies. J.Amer.Statist. Assoc. 62 1121-1139.

Seely, J.F. (1979) Large sample properties of invariant quadratic unbiased estimators in the random one-way model. In Van Vleck and Searle (1979).

Shah, B.V. (1978) Variance estimates for complex statistics from multistage sample surveys. In Namboodiri (1978).

Shah, B.V., Holt, M.M. and Folson, R.E. (1977) Inference about regression models from sample survey data. Bull.Int. Statist. Inst. Book 3 43-57.

Shuster, J.J. and Downing, D.J. (1976) Two-way contingency tables for complex sampling schemes. Biometrika 63 271-276.

Sibson, R. (1979) Studies in the robustness of multidimensional scaling: perturbation analysis of classical scaling. J.Roy.Statist. Soc. B 41 217-229.

Skinner, C.J. (1981) Estimation of the variance of a finite population for cluster samples. Sankhya B 392-398.

Skinner, C.J. (1982) Multivariate prediction from selected samples. Biometrika. Accepted for publication.

Smith, H.F. (1938) An empirical law describing heterogeneity in the yields of agriucltural crops. J.Agric.Sci. 28 1-23.

Smith, T.M.F. (1976) The foundations of survey sampling. J.Roy. Statist. Soc. A. 139 183-195.

Smith, T.M.F. (1978) A model building approach to survey analysis. European Meeting of Statisticians, Oslo, August, 1978.

Smith, T.M.F. (1982) Regression analysis for complex surveys. In 'Current Topics in Survey Sampling. D. Krewski, J.N.K. Rao and R. Platek (Eds) Academic Press, New York.

Sugden, R.A. (1980) Partial exchangeability and inference on stratified populations. Manuscript.

Sugiura, N. (1976) Asymptotic expansions of the distributions of the latent roots and the latent vectors of the Wishart and multivriate F matrices. J. Mult.Anal. 6 500-525.

Swallow, W.H. (1981) Variances of locally minimum variance quadratic unbiased estimators (MIVQUES) of variance components. Technometrics 23 271-283.

Tan, W.Y. (1978) On the quadratic estimation of covariance matrices in MANOVA random effects models. Statistica 38 449-458.

Tan, W.Y. (1979) On the quadratic estimation of covariance matrices in multivariate linear models. J.Mult.Anal. 9 452-459.

Thompson, R. (1977) Discussion of Dempster et al (1977) p.34.

Thomsen, I. (1978) Design and Estimation problems when estimating a regression coefficient from survey data. Metrika 25 27-35.

Thomson, G.H. (1938) The influence of univariate selection on the factorial analysis of ability. Brit.J.Psych. 28 451-459.

Thomson, G.H. (1951) The Factorial Analysis of Human Ability 5th Ed. Univ. of London Press, London.

Thomson, G.H. and Ledermann, W. (1939) The influence of multivariate selection on the factorial analysis of ability. Brit.J.Psych. 29 288-305.

Thurstone, L.L. (1945) The effects of selection in factor analysis. Psychometrika 10 165-198.

Tomberlin, T.J. (1979) The analysis of contingency tables of data from complex samples. Amer.Statist.Assoc. Proc. of Sect. on Survey Research Methods 152-157.

Tomberlin, T.J. (1980) A model-based approach to the analysis of contingency tables of data from complex samples. Amer.Statist.Assoc. Proc. of Sect. on Survey Research Methods. 230-234.

Tortora, R.D. (1980) The effect of disproportionate stratified design on principal components analysis used for variable elimination. Amer.Statist.Assoc.Proc. of Sect. on Survey Research Methods 746-750.

Tyler, D.E. (1981) Asymptotic inference for eigenvectors.Ann.Statist. 9 725-736.

Van Vleck, L.D. and Searle, S.R. (Eds) (1979) Variance Components and Animal Breeding. Cornell Univ. Ithaca, New York.

Wakimoto, K. (1971a) Stratified random sampling (I); estimation of the population variance. Ann.Inst.Statist.Math. 23 233-252.

Wakimoto, K. (1971b) Stratified random sampling (II); estimation of the population covariance. Ann.Inst.Statist.Math. 23 327-337.

Wakimoto, K. (1971c) Stratified random sampling (III); estimation of the population correlation coefficient. Ann.Inst.Statist.Math. 23 339-353.

Walsh, J.E. (1947) Concerning the effect of intra-class correlation on certain significance tests. Ann.Math.Statist. 18 88-96.

Waternaux, C.M. (1976) Asymptotic distribution of the sample roots for a nonnormal population. Biometrika 63 639-645.

Whittle, P. (1956) On the variation of yield variance with plot size. Biometrika 43 337-343.

Whittle, P. (1962) Topographic correlation, power-law covariance functions and diffusion. Biometrika 49 305-314.

Wilkinson, J.H. (1965) The Algebraic Eigenvalue Problem. Clarendon Press, Oxford.

Woodruff, R.S. (1971) A simple method for approximating the variance of a complicated estimate. J.Amer.Statist.Assoc. 66 411-414.

Yates, F. (1960) Sampling Methods for Censuses and Surveys 3rd Ed. Griffin, London.

Zacks, S. (1981) Bayes equivariant estimators of the variance of a finite population for exponential priors. Comm. Statist. A10 427-437.

Zacks, S. and Solomon, H. (1981) Bayes and equivariant estimators of the variance of a finite population: Part I, simple random sampling. Comm.Statist. A10 407-426.