TOPICS IN STATISTICAL DISCRIMINATION

by

OMAR MOHD. RIJAL

A dissertation submitted to the

UNIVERSITY OF GLASGOW

for the degree of

Doctor of Philosophy

1984

Department of Statistics, December 1984.

ACKNOWLEDGEMENTS

I would like to express my thanks to my supervisors Dr. I. Ford and Mr. A.D. McLaren for their advice and encouragement during the period of this research. I would also like to thank Dr. F. Critchley for his valuable contributions, and to all of the members of the Statistics Department of Glasgow University for their willingness to lend a sympathetic ear to my problems during my postgraduate career. Finally, I would like to thank Mrs. M. Smith for her patient typing of this thesis.

TABLE OF CONTENTS

The relevant Appendices are placed at the end of either PART(I) or PART(II). Within each PART, the Appendices are numbered to correspond with the appropriate Chapter. References for both parts are given at the very end.

TABLE OF CONTENTS (contd.)

SUMMARY

This thesis is in two distinct sections, PART(I) and PART(II). The study in PART(I) is independent of that in PART(II).

PART(I)

We look at a particular data set in PART(I) where seven observers use four output media to discriminate positive radiological images from null ones. The main task is to rank the output media in order of effectiveness. Each observer makes one of five possible responses, but, for the most part, we simplify the problem by treating it as one of binary response.

In Chapter 1 we use a technique of decision theory. The criterion for preferring one of two output media involves comparison of the probabilities of deciding a null image is positive and vice versa.

In Chapter 2, using linear models, in particular logistic linear models, we look for any possible interaction between observers and output media and, if possible, rank the latter.

Chapter 3, collects the results of the previous two chapters. An analogy between the criteria used in Chapters 1 and 2 is given geometrically.

PART(II)

In this section we study the two population discrimination problem for the particular case of multivariate normal data with equal or unequal covariance matrices. We study the estimation of the log-odds, $\theta(\underline{x})$, and the use of approximate interval estimates for $\theta(\underline{x})$ as a means of expressing uncertainty due to the estimation involved.

An immediate difficulty when the covariance matrices are unequal is that it is not possible to derive the exact variance of the estimated log-odds, $\hat{\theta}(\underline{x})$. In Chapter 2 an approximate variance for $\hat{\theta}(\underline{x})$ is

derived. Chapter 2 contains other technical results that we use in later chapters.

Chapter 3 is an extensive simulation study. This involves the study of

(i) the empirical distribution of $\hat{\theta}(\underline{x})$

(ii) the performance of approximate interval estimation methods for $\hat{\theta}(\underline{x})$, assuming equal covariances, unequal covariances and when a test of equality of covariances is carried out to decide which method to use.

We study the effects, on (i) and (ii) of varying various parameters. These parameters include sample sizes, dimensionality and various forms for the true covariance structure.

We apply the approximate interval estimation methods to several data sets in Chapter 4. Informative plots are given as a pictorial aid in solving the discrimination problem. These plots are useful for the study of misclassification properties and outliers, as well as for the classification of new cases.

The results of both Chapters 3 and 4 indicate that for small sample sizes and/or large dimensionality we have problems in constructing useful interval estimates for $\theta(\underline{x})$. In Chapter 5 we study alternative methods of interval estimation for $\theta(\underline{x})$ with emphasis on small sample situations. The methods involve,

(i) Alternative variance estimates

(ii) Bootstrapping, Efron (1981, 1982)

(iii) Profile-likelihood, Kalbfleisch (1979).

Part(II) ends with a brief look at the effect of the non-normality of the distribution of $\underline{x}$ on the approximate methods. This involves a brief simulation study investigating the effect of increasing skewness and kurtosis of the distribution of $\underline{x}$ on the results.

CHAPTER 1: DECISION RULES FOR NUCLEAR IMAGING FORMATS

(1.1) THE EXPERIMENT

(A) Introduction

Analogue signals direct from a Gamma camera have been used to output images onto either Polaroid or single sided X-ray film (to be referred to as POLAROID and ANALOGUE X-ray respectively). Digital images stored by the computer on 'floppy' discs could be reproduced on either polaroid or X-ray film. A television camera system, 'Vidicam, Tudorcape Ltd', was used to output the image on the visual display unit of the image processor onto half-plate black and white photographs. (The second pair of hard copy outputs are to be referred to as Digital X-ray and Vidicam respectively).

One hundred images were produced on each of the four selected output media, seven observers looked at the four sets of 100 images. In 50 of these 100 images an absorber was used to produce cold spots of 5 different visibilities. These cold spots are 'shades' or 'spots' produced on the hard copy output.

Each observer was asked to place each image into one of five categories of confidence, namely:-

B_1 = cold spot definitely present

B_2 = cold spot probably present

B_3 = cold spot possibly present

B_4 = cold spot probably not present

B_5 = cold spot definitely not present.

The B_i's could be considered in the following combinations, to produce a definite decision rule.

	Decide cold spot present if we say
Rule 1	B_1
Rule 2	B_1 or B_2
Rule 3	B_1 or B_2 or B_3
Rule 4	B_1 or B_2 or B_3 or B_4
Rule 5	all B_i

TABLE [1(I)]

The main objective of this study is to compare the four different output media.

(B) R.O.C. Analysis

The results for the experiment in Section 1.1(A) were obtained from Eadie et al (1980), and tabulated as shown in Appendix (1.1). Analysis of the results was done by Eadie et al (1980) using a technique called R.O.C. analysis [Metz (1978), Lusted (1971, 1978)]. We will now briefly illustrate this technique, described in terms of the experiment done by Eadie et al (1980).

Firstly, films or 'pictures' containing a cold spot are classified as POSITIVE CASES. Fifty of the films do not have cold spots and are classified as NEGATIVE CASES. Using definitions from Metz (1978), and the 'set-up' of Table [1(II)], we have

$$P(\text{true positive}) \equiv P(T) \equiv \text{Prob}\{\text{decide cold spot present} \mid \text{POSITIVE CASE}\}$$

$$P(\text{false positive} \equiv P(F) \equiv \text{Prob}\{\text{decide cold spot present} \mid \text{NEGATIVE CASE}\}$$

Further, referring to Table [1(I)] and Table [1(II)], RULE 1 to RULE 5 clearly gives us five pairs of P(T) and P(F). The plot of these

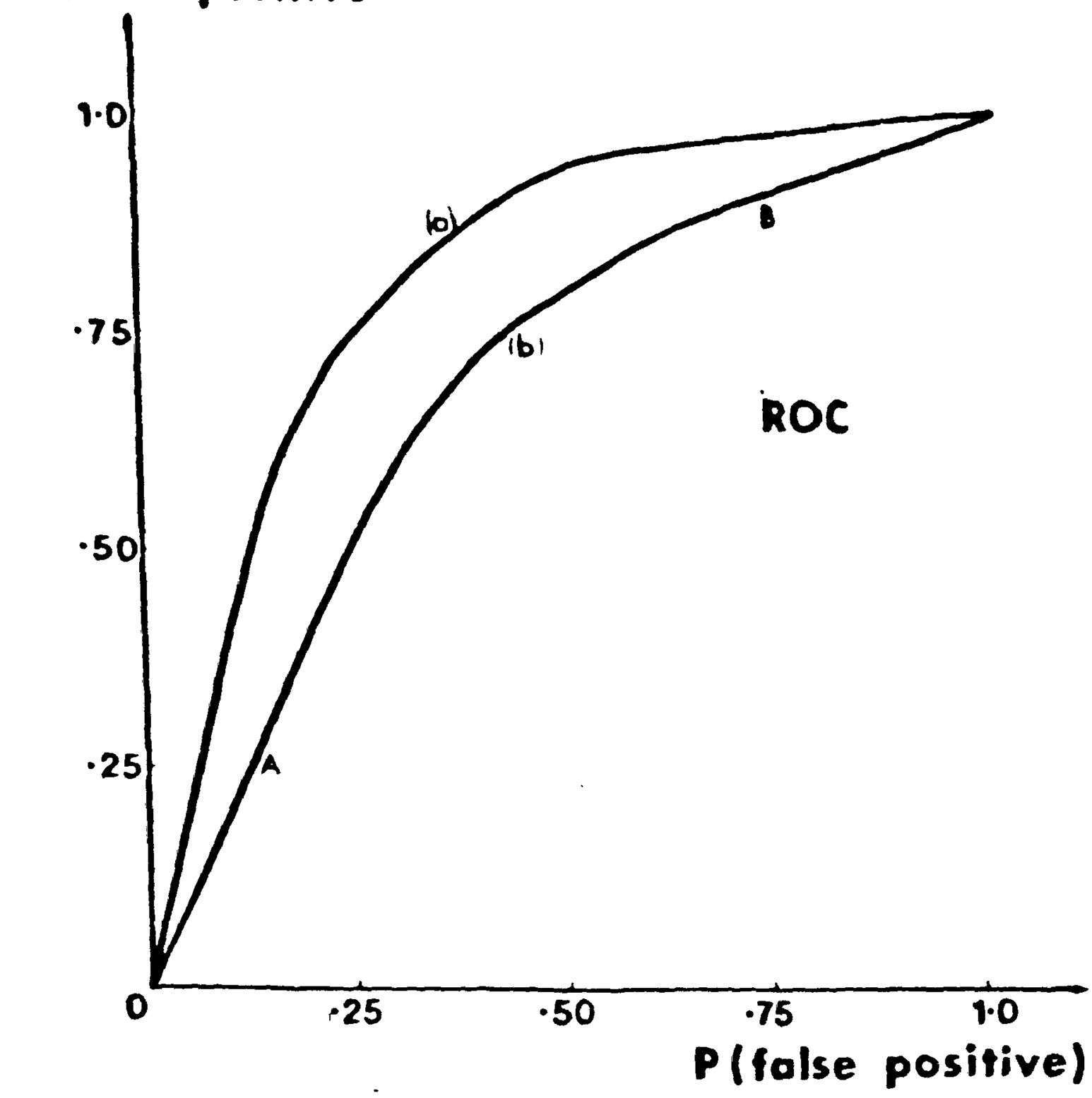
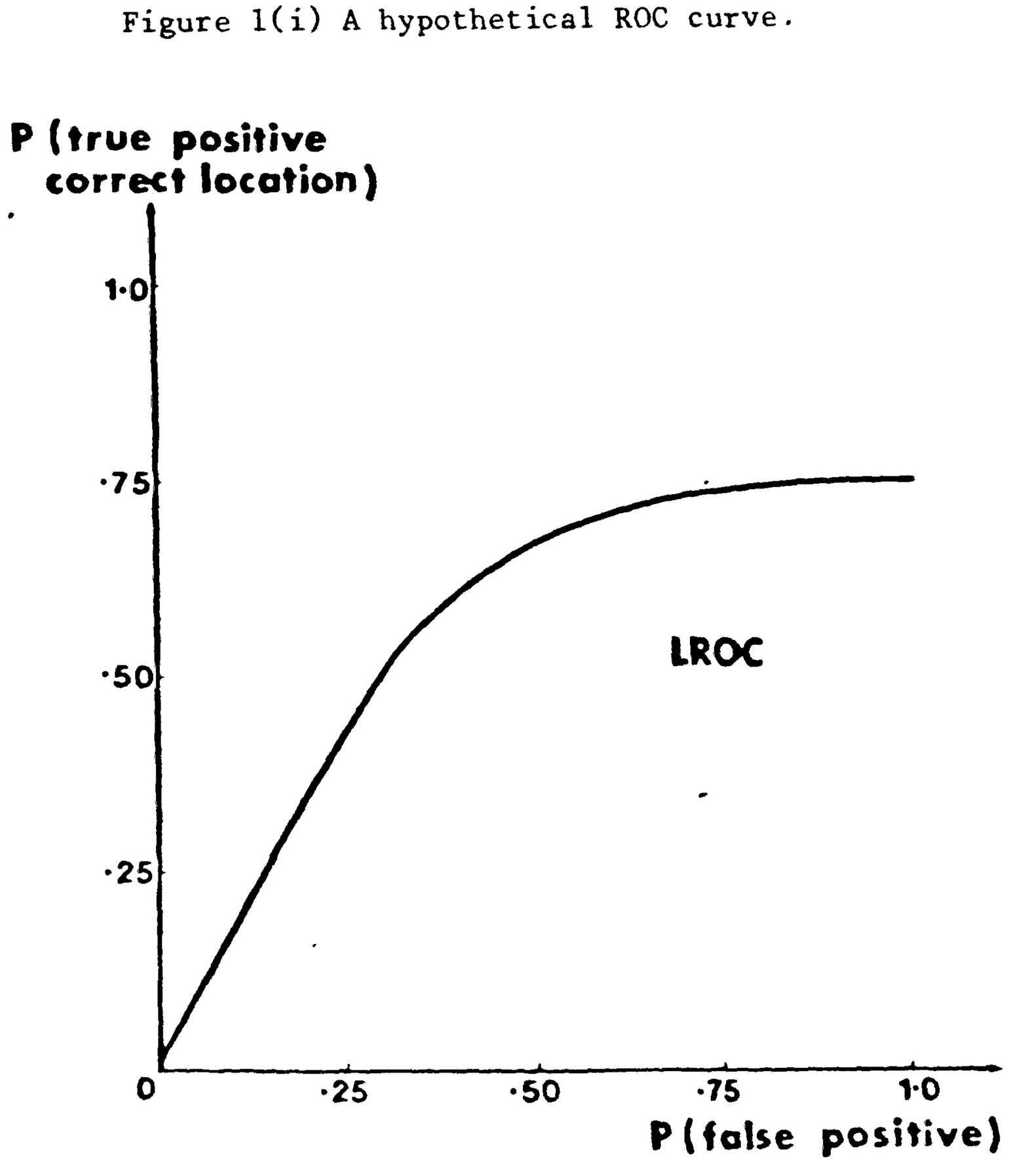


Figure 1(i) A hypothetical ROC curve.



A hypothetical LROC curve.

five points lies on a hypothetical R.O.C. curve, for example see figure [1(I)]. In figure [1(I)], point A could be a clinical situation where a cautious decision (low-risk of false positive) is made; and point B could be a situation where the decision maker associates a 'small' cost (or loss) involved in making a false positive decision. This is essentially the main feature of using the R.O.C. curve. Further, curve (a) is 'said' to be better than curve (b), in figure [1(I)], because for a given P(T) the latter curve has a larger P(F).

If we replace P(T), in figure [1(I)] by

$$\text{Prob}\left(\begin{array}{l}\text{true}\\ \text{(positive)}\\ \text{(correct)}\\ \text{(location)}\end{array}\right) = \text{Prob}\left(\begin{array}{l}\text{Decide cold spot present}\\ \text{(and correctly locate cold spot}\end{array}\middle|\begin{array}{l}\text{POSITIVE}\\ \text{CASE}\end{array}\right)$$

we have the LROC curve in figure [1(II)] [see for example Houston and Macleod (1979)]

(1.2) AN ESTIMATED LIKELIHOOD RATIO RULE

For a given observer and output medium,

let m_i = number of response B_i made given POSITIVE cases

n_i = " " " " NEGATIVE "

			OBSERVER	
			POSITIVE CASES	NEGATIVE CASES
TYPE OF OUTPUT MEDIUM	RESPONSE	B_1	m_1	n_1
	"	B_2	m_2	n_2
	"	B_3	m_3	n_3
	"	B_4	m_4	n_4
	"	B_5	m_5	n_5

TABLE [1(ii)]

where $\sum_{i=1}^{5} m_i = 50 = \sum_{i=1}^{5} n_i$,

and B_1. B_2..... B_5 are defined to be mutually exclusive and exhaustive categories.

We proceed to establish a decision rule. and in doing so investigate the 'validity' of RULE 1 to RULE 5 as defined in Table [1(i)]. This 'validity' of RULE 1 to RULE 5 will partially support the use of the R.O.C. points (P(T), P(F)).

$$\text{Define } \theta_i = \text{prob}\left(\begin{matrix}\text{Response} \\ B_i\end{matrix}\middle|\ \text{POSITIVE CASE}\right) \qquad \phi_i = \text{prob}\left(\quad " \quad\middle|\ \text{NEGATIVE CASE}\right)$$

for given Observer/output medium

$$\text{where } \sum_{i=1}^{5} \theta_i = 1 = \sum_{i=1}^{5} \phi_i$$

we have probability distributions of $(m_1, \ldots, m_5)$ and $(n_1, \ldots, n_5)$ as Multinomials. viz:-

$$p(m_1, \ldots, m_5) = \frac{50!}{m_1!m_2!\ldots m_5!}\ \theta_1^{m_1}\ \theta_2^{m_2} \ldots\ \theta_5^{m_5}$$

$$p(n_1, \ldots, n_5) = \frac{50!}{n_1!n_2!\ldots n_5!}\ \phi_1^{n_1}\ \phi_2^{n_2} \ldots\ \phi_5^{n_5}$$

Hence we have

		OBSERVER	
		POSITIVE CASE	NEGATIVE CASE
OUTPUT MEDIUM	B_1	θ_1	ϕ_1
	B_2	θ_2	.
	B_3	.	.
	B_4	.	.
	B_5	θ_5	ϕ_5

TABLE [1(iii)]

Maximum Likelihood estimates of θ_i. ϕ_i are

$\hat{\theta}_i = \frac{m_i}{50}$ and $\hat{\phi}_i = \frac{n_i}{50}$ (result (13.4.18) of Bishop et al (1975))

and it is these that appear in Appendix (1.1). The data in Appendix (1.1) was obtained from Eadie et al (1980).

What we would like now is to consider the situation:- suppose any of 7 observers makes an observation (his 101^{th} observation) and makes response B_i. Can we tell from this response (or observation) that the cold spot is present?

Hence we have two decisions between which to choose: decide cold spot present or absent. Each decision depends on whether observation B_i is made from the sample of POSITIVE CASES or NEGATIVE CASES.

The likelihood ratio rule then takes the form.

Decide cold spot present if $\lambda(B_i) > C$

or Decide cold spot absent if $\lambda(B_i) < C$

and Decide "present" or "absent" if $\lambda(B_i) = C$

where $0 < C < \infty$

and $\lambda(B_i) = \frac{\theta_i}{\Phi_i}$

We use the estimate $\hat{\lambda}(B_i) = \frac{\hat{\theta}_i}{\hat{\Phi}_i}$

Let us denote the Likelihood Ratio rule as LR(C). What we discovered in most cases was that $\frac{\hat{\theta}_i}{\hat{\Phi}_i} > \frac{\hat{\theta}_{i+1}}{\hat{\Phi}_{i+1}}$, $i = 1,2,3,4$

(See Appendix (1.2))

so $\frac{\theta_i}{\hat{\Phi}_i} > C$ would mean "Decide cold spot present" if Response is

B_i or B_{i-1} or B_{i-2} ... . We illustrate by an example.

Suppose $\frac{\hat{\theta}_3}{\hat{\Phi}_3} < C < \frac{\hat{\theta}_2}{\hat{\Phi}_2}$

then, $\frac{\hat{\theta}_1}{\hat{\Phi}_1}$, $\frac{\hat{\theta}_2}{\hat{\Phi}_2}$ are greater than C and

$\frac{\hat{\theta}_3}{\hat{\Phi}_3}$, $\frac{\hat{\theta}_4}{\hat{\Phi}_4}$, $\frac{\hat{\theta}_5}{\hat{\Phi}_5}$ are less than C.

and by an estimated likelihood Ratio Rule we say,

decide cold spot present if Response is B_1 or B_2

" " " absent " " " B_3, B_4 or B_5.

Of course depending on what values C may take, we can use 4 decision rules. These rules are in fact estimated Likelihood Ratio Rules, and correspond to the rules listed in Table (1(i)). We can now say that the R.O.C. points (P(T), P(F)) has a 'sensible' interpretation.

The current emphasis on Likelihood ratio decision rules is that these rules have several optimal properties which we briefly mention.

Firstly we have two types of errors,

α = Prob(Decide cold spot absent | Positive case)

β = Prob(Decide cold spot present | Negative case)

Theorem

A rule LR(C) minimises $\alpha + C\beta$ among all possible decision rules based on the observation B_i. (See Appendix (1.3))

Corollary

Let α, β be error probabilities for LR(C)

Let α^*, β^* be error probabilities for another decision rule, then

(i) if $\alpha^* < \alpha$ then $\beta < \beta^*$ [have to be true for Theorem, or for $(\alpha + C\beta > \alpha^* + C\beta^*)$ to be true]

(ii) if $\beta^* < \beta$ then $\alpha < \alpha^*$ [same reason as (i)].

We note that this corollary is essentially the Neyman Pearson Lemma.

(1.3) Application to Eadie's Data

(A) A comparison Criterion

Keeping closely at this stage to the R.O.C. analysis done by Eadie et al (1980) we recall some earlier definitions.

P(T) = Prob(decide cold spot present|Positive CASE)

P(F) = Prob(decide cold spot present|Negative CASE)

Hence α = Prob(Decide cold spot absent|Positive CASE) = 1 - P(T)

and β = Prob(Decide cold spot present|Negative CASE) = P(F)

Using Likelihood Ratio rules LR(C) minimises $\alpha + C\beta$ or minimises [1 - P(T) + C P(F)]

$\equiv$ maximise [P(T) - C P(F) - 1]

$\equiv$ maximise [P(T) - C P(F)]

The value of C reflects the importance of the (α, β) errors (namely, is making α-error more serious than making β-error). Discussions with Eadie et al (1980) have led to the use of the value C = 1.

Therefore Minimising $\alpha + C\beta$ is equivalent to Maximising P(T) - C P(F) [with C = 1 by choice]. It now leaves to express P(T) and P(F) in terms of the θ_i's and ϕ_i's.

Firstly

$$P(T) = \text{Prob}\begin{pmatrix}\text{decide cold} \\ \text{spot present}\end{pmatrix}\Big|\begin{matrix}\text{POSITIVE} \\ \text{CASE}\end{matrix}\Big)$$

$$= \begin{cases} \text{Prob}\begin{pmatrix}\text{decide } B_1 \\ \text{or } B_2\end{pmatrix}\Big|\begin{matrix}\text{POSITIVE} \\ \text{CASE}\end{matrix}\Big) = \theta_1 + \theta_2 \rightarrow \text{RULE 2} \\ \\ \text{Prob}\begin{pmatrix}\text{decide } B_1 \\ \text{or } B_2 \text{ or } B_3\end{pmatrix}\Big|\begin{matrix}\text{POSITIVE} \\ \text{CASE}\end{matrix}\Big) = \theta_1 + \theta_2 + \theta_3 \rightarrow \text{RULE 3} \end{cases}$$

Similarly

$$P(F) = \begin{cases} \text{Prob}\begin{pmatrix}\text{decide } B_1 \\ \text{or } B_2\end{pmatrix}\Big|\begin{matrix}\text{NEGATIVE} \\ \text{CASE}\end{matrix}\Big) = \phi_1 + \phi_2 \quad \text{(RULE 2)} \\ \\ \text{Prob}\begin{pmatrix}\text{decide } B_1 \\ \text{or } B_2 \text{ or } B_3\end{pmatrix}\Big|\begin{matrix}\text{NEGATIVE} \\ \text{CASE}\end{matrix}\Big) = \phi_1 + \phi_2 + \phi_3 \quad \text{(RULE 3)} \end{cases}$$

Suppose we now compare the output media Vidicam and Polaroid (for given observer)

Define $D_{IJ} = [P(T) - P(F)]_I - [P(T) - P(F)]_J$

$= D_I - D_J$

where I $\equiv$ Vidicam and J $\equiv$ Polaroid.

Since we are trying to maximise P(T) - P(F) we would like D_I and D_J to be both positive. But if D_{IJ} is positive, D_I should be more positive than D_J, which means that using Vidicam minimises the α, β errors more than Polaroid. In this sense Vidicam is said to be the 'better' of the two [for a given rule].

We still need to show that D_I is significantly different from D_J i.e. ($D_I \neq D_J$). This is done by finding the approximate interval estimates: $\hat{D}_{IJ} \pm 1.96 \sqrt{\widehat{\text{variance}}(\hat{D}_{IJ})}$ (see Appendix (1.4)) where the symbol '$\wedge$' denotes estimates.

(B) Some results

No significant results were obtained for Rule 1 and Rule 2. For Rule 3 we have only two conclusive results, observer 6 did better on Vidicam as against Digital and Polaroid (see Table [1(iv)]). In Table [1(iv)] only the interval estimates of D_{IJ} that exclude zero are listed.

For Rule 4, most of the upper limits of the interval are close to zero. Note also that for Rule 4 and observer 7, the value of $\hat{D}_I$ is negative. This suggests that the likelihood ratio rules, as defined in Section 1.1(C), may not be valid in some cases.

Table [1(iv)] also suggests (or shows)

(i) Vidicam best output medium

(ii) Observer 5 cannot distinguish between any two media, no significant interval estimates for RULE 4.

(iii) Different observers perform differently on the same or different output media for given rule.

As most of the results in Table [1(iv)] are for rule 4, we recall its definition from Table [1(i)]. An observer applying rule 4 is one who is anything from being very sure to possibly guessing when making his decisions. Despite this 'relaxed' criterion, less than half

of the total observer and output combinations gave significant interval estimates. Clearly the interval estimates of D_{IJ} may not be 'sensitive' enough to pick out differences between output media. Perhaps we should try a totally different approach to the original problem of deciding which output medium is better, and in doing so, verify some of our current results.

In view of (iii) we proceed with the question "Is there interaction between observer and output media for given rule?". Interaction is taken as 'standard' terminology in linear regression techniques. We consider interaction for a given rule in the belief that differentiating between B_2, B_3 and B_4 [Table 1(i)] is difficult, perhaps unwise. We therefore ignore the possibility of different rules (as defined in Table 1(i)) for different observer/output combinations.

I	J	$\hat{D}_I$	$\hat{D}_J$	$\hat{D}_{IJ}$	$\hat{D}_{IJ} \pm 1.96 \sqrt{\hat{Var}(\hat{D}_{IJ})}$	RULE/ OBSERVER
1	4#	0.08	0.54	-0.46	-0.71, -0.21	3, 6
3	4#	0.26	0.54	-0.28	-0.53, -0.03	3, 6
1	2#	0.20	0.44	-0.24	-0.47, -0.01	4, 1
2#	3	0.44	0.12	0.32	0.13, 0.51	4, 1
3	4#	0.12	0.38	-0.26	-0.44, -0.08	4, 1
1	2#	0.06	0.36	-0.30	-0.55, -0.05	4, 2
1	3#	0.06	0.36	-0.30	-0.54, -0.06	4, 2
1	2#	0.12	0.42	-0.30	-0.53, -0.07	4, 3
1	4#	0.12	0.42	-0.30	-0.54, -0.06	4, 3
2#	3	0.42	0.08	0.34	0.13, 0.55	4, 3
3	4#	0.08	0.42	-0.34	-0.57, -0.11	4, 3
1	4#	0.12	0.42	-0.30	-0.56, -0.04	4, 4
1	4#	0.00	0.30	-0.30	-0.50, -0.10	4, 6
2	4#	0.06	0.30	-0.24	-0.42, -0.06	4, 6
3	4#	0.08	0.30	-0.22	-0.42, -0.02	4, 6
1	2#	-0.04	0.26	-0.30	-0.54, -0.06	4, 7
1	3#	-0.04	0.22	-0.26	-0.47, -0.05	4, 7

TABLE [1(iv)]: Significant interval estimates of D_{IJ}

where (i) $D_{IJ} \approx D_I - D_J$

$D_I \approx [P(T) - P(F)]_I$

(ii) {I} or {J} ≡ {1, 2, 3, 4} ≡ {Digital, Analogue, Polaroid, Vidicam} respectively.

(iii) See Table (1(i)) for definition of RULE and Appendix (1.1) for definition of observer-labels.

(iv) Symbol # denotes "choose this output medium".

CHAPTER 2: USING LINEAR MODELS

(2.1) A LOG-LINEAR MODEL

In defining P(T) and P(F) previously, we have counts of decisions (decide cold spot present or absent) for given observer and output medium. We now use these counts to formulate linear models with the purpose of detecting (in particular) observer and output medium interaction. The TRUE and FALSE POSITIVES will be separately considered for the moment.

Let X_{ijk} = observed number of decision i, given output medium j, and observer k [i=1 if decide cold spot present, i=2 if decide cold spot absent, j=1,2,3,4, and k = 1,2,3,...,7].

The data in Appendix (1.1) [i.e. Eadie's data] can be retabulated such that $X_{1jk} \sim Bin(50, Px)$

given $X_{1jk} + X_{2jk} = 50 \ \forall$ j,k

and $\sum_i \sum_j \sum_k X_{ijk} = N = 7 \times 4 \times 50$

Given X_{1jk}, $X_{2jk}(\forall j,k)$ are independent non-negative integer random variables and

given $P[X_{1jk} = x_{1jk} | X_{1jk} + X_{2jk} = x_{.jk}]$ (2.1.1)

$\sim Bin(x_{.jk}, Px)$

$(x_{1jk} = 0,1,2..,x_{.jk})$

Then we have from Chatterji (1963) $X_{1jk} \sim P_o(\theta_{1jk})$, $X_{2jk} \sim P_o(\theta_{2jk})$ with

$$\frac{\theta_{1jk}}{\theta_{2jk}} = \frac{Px}{1-Px}$$

where $P_o(\theta)$ denote the Poisson distribution with mean θ and Bin(n,p) denote the binomial distribution with parameters n,p.

(2.1.1) can be restated as,

$$P(X_{1jk} | x_{.jk}) \sim Bin\left[x_{.jk}, \frac{\theta_{1jk}}{\theta_{1jk} + \theta_{2jk}}\right] \qquad (2.1.2)$$

In particular,

$$P(X_{ijk}) = P(X_{.jk})P(X_{ijk}|x_{.jk}) \qquad (2.1.3)$$

we now explain how our binomial data can be analysed as if all of the counts were independent Poisson random variables. The 'usual' hypothesis of independence is stated as,

$$\theta_{ijk} = m\Pi_i\Pi_j\Pi_k \qquad (2.1.4)$$

where Π_k = probability of observer k making a decision

Π_j = probability output medium j used

Π_i = probability decision i made

m = constant

and $\sum_i \Pi_i = \sum_j \Pi_j = \sum_k \Pi_k = 1$

Let $\Phi_{jk} = \theta_{1jk}/(\theta_{1jk} + \theta_{2jk})$

By (2.1.4), $\Phi_{jk} = \dfrac{\Pi_1}{\Pi_1 + \Pi_2}$ which does not

depend on observer and output media. The hypothesis in (2.1.4) can be restated in the following form.

$$\log(\theta_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k \qquad (2.1.5)$$

where μ, α_i, β_j, γ_k are the logarithms of m, Π_i, Π_j, Π_k respectively.

The 'log-linear' model given in (2.1.5) implies,

$$\Phi_{jk} = \frac{\exp(\alpha_1)}{[\exp(\alpha_1) + \exp(\alpha_2)]}$$

that is, model (2.1.5) is equivalent to the probability of "decide cold spot present" is independent of observer or output medium.

The hypothesis of "interaction" between, say making decision i, and using output medium j can be stated as,

$$\log(\theta_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} \qquad (2.1.6)$$

where $(\alpha\beta)_{ij}$ is called the interaction term. Model (2.1.6) implies,

$$\Phi_{jk} = \frac{\exp[\alpha_1 + (\alpha\beta)_{1j}]}{\exp[\alpha_1 + (\alpha\beta)_{1j}] + \exp[\alpha_2 + (\alpha\beta)_{2j}]}$$

i.e. the probability of decision 1 depends only on output medium j.

Thus interaction in the Poisson model corresponds to the main effect in the actual binomial model.

The 'main effects' terms α_i, β_j, γ_k and the 'grand mean' μ from model (2.1.4) can be expressed in terms of the expected counts θ_{ijk}. For example, Bishop et al (1975) show that

$$\mu = \frac{1}{IJK} \sum_i \sum_j \sum_k \log(\theta_{ijk})$$

$$\alpha_i = \frac{1}{JK} \sum_j \sum_k \log(\theta_{ijk}) - \mu$$

$$\beta_j = \frac{1}{IK} \sum_i \sum_k \log(\theta_{ijk}) - \mu$$

etc...

The values of the 'main effects' and interaction terms is clearly affected by the design of the experiment, viz:- $x_{.jk} = 50$.

For interpretation purposes we will consider hierarchical models as defined by Bishop et al (1975).

An alternative way of looking at these models is that since X_{ijk} is independently distributed as $P_o(\theta_{ijk})$, then $P(X_{ijk} | \sum_{ijk} X_{ijk} = N)$ (for all i,j,k) is multinomial. We can think of X_{ijk} as observed counts in a Contingency Table.

2.2: An algorithm for estimating the parameters

We estimate the parameters of the log-linear model by making use of the computer Package 'GLIM' (Baker and Nelder (1978)). GLIM makes use of an iterative Newton Raphson method to find the maximum Likelihood estimate of the parameters. Nelder and Wedderburn (1972) shows that the likelihood estimates are unique for the log-linear model. A detailed explanation of GLIM is given in McCullagh and Nelder (1983) where they showed, for example, that the Newton-Raphson method used is equivalent to a weighted least squares method.

The goodness of fit for the models is based on,

$$\lambda = \frac{\max \text{Lik}\,[\theta_{ijk};\ X_{ijk}] \quad \text{(full model)}}{\max \text{Lik}\,[\theta_{ijk};\ X_{ijk}] \quad \text{(current model)}}$$

where Lik[.] denotes a likelihood function. The current model [or null hypothesis] could be, say (2.1.5), and the full model (Alternative hypothesis) is the log-linear model with the $(\alpha\beta\gamma)_{ijk}$ term. It is a well-known result, e.g. Bishop et al (1975), Chapter 4, that 2 log λ [the DEVIANCE] is asymptotically chi-squared with K degrees of freedom, where

K = (dimension of full model)

- (dimension of current model).

The approximation is good if, X_{ijk} is large and the hypothesis for 'current' model is true. The models tested are given in Table 2(i), together with their interpretation.

We note that the ratio λ is a ratio of maximised likelihoods of two Poisson distributions corresponding to two hypotheses. In Appendix (2.2) we show that this ratio is equivalent to the corresponding ratio of maximised likelihoods for the binomial, which is the situation we are looking at.

2.3 Model fitting to Eadie's (1980) data

(A) Looking for a structure

We now fit the models as given in Table 2(i) to the data (essentially Appendix 1.1) such that for a given RULE, OBSERVER and OUTPUT MEDIUM we have two responses (i.e. decide cold spot present or otherwise). In Table [2(ii)] are the DEVIANCE and 'degrees of freedom' for the four models.

For the False positives, none of the models fit the data for RULE 3 and RULE 4. We accept model 3 using RULE 2.

For the True positives, again RULE 4 rejects all models. RULE 2 and RULE 3 accepts model 1, while RULE 1 accepts model 3.

MODEL	INTERPRETATION
1. $\ln(\theta_{ijk}) = \mu+\alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk}$	Decision* is completely independent of observer and output medium.
2. $\ln(\theta_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + (\alpha\beta)_{ij}$	Decision* depends only on output medium j.
3. $\ln(\theta_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik}$	Decision* depends only on observer k.
4. $\ln(\theta_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + (\alpha\beta)_{ij}$	Decision* depends on both output medium and observer, whose effects combine additively on the logistic scale.

TABLE 2(i): Models considered

Note 1: (*): Decision means the conditional probability of decide cold spot present given output medium and observer, see (2.1.2)

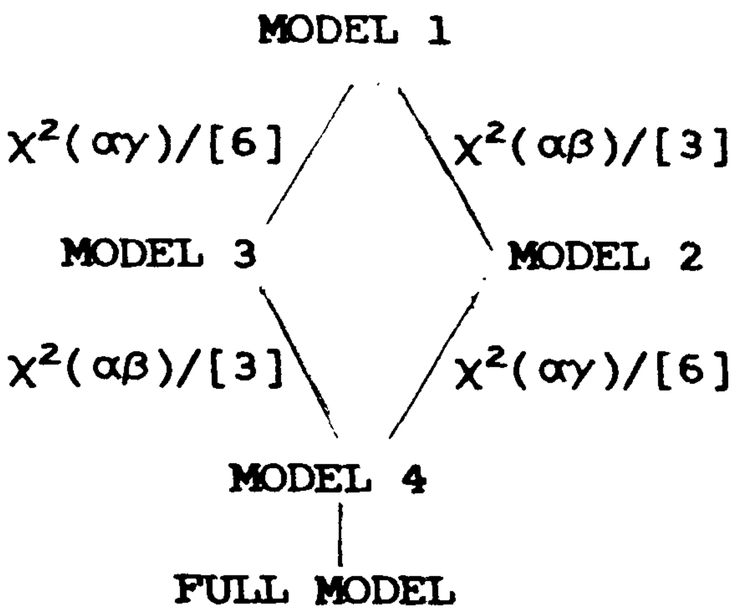
Note 2: In Model (1), the term $(\beta\gamma)_{jk}$ is included because $\sum_i X_{ijk}$ is fixed, see for e.g. Everitt (1977) section (5.4).

	MODEL	RULE 1	RULE 2	RULE 3	RULE 4
TRUE POSITIVES	1	51.48 (27)	21.93 (27)	34.63 (27)	158.70 (27)
	2	49.22 (24)	21.70 (24)	33.23 (24)	146.80 (24)
	3	8.97 (21)	8.28 (21)	13.13 (21)	42.91 (21)
	4	6.64 (18)	8.05 (18)	11.71 (18)	29.73 (18)
FALSE POSITIVES	1	#	72.85 (27)	221.10 (27)	373.4 (27)
	2	#	66.92 (24)	207.10 (24)	339.3 (24)
	3	#	24.25 (21)	52.68 (21)	100.5 (21)
	4	#	18.09 (18)	35.56 (18)	58.81 (18)

TABLE [2(ii)]: In each cell, the pair of numbers are the Deviance and degrees of freedom (latter in parentheses)

(#) Note: Too many cell counts are zero

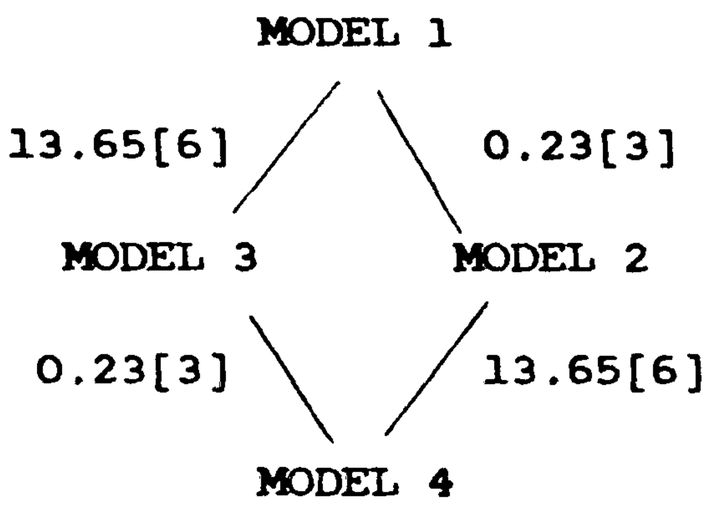
RULE 4 suggests 'no simple structure' in the models used. For RULE 3 this is again true for the false positives, in contrast to accepting the 'simplest' model for the true positives. RULE 2 is somewhat similar to RULE 3. We further note that the values of the DEVIANCE tend to be smaller for the TRUE POSITIVES. These remarks suggest investigating the TRUE POSITIVES in more detail, in particular the difference in Deviances between two models. We illustrate, for RULE 2 and RULE 3.



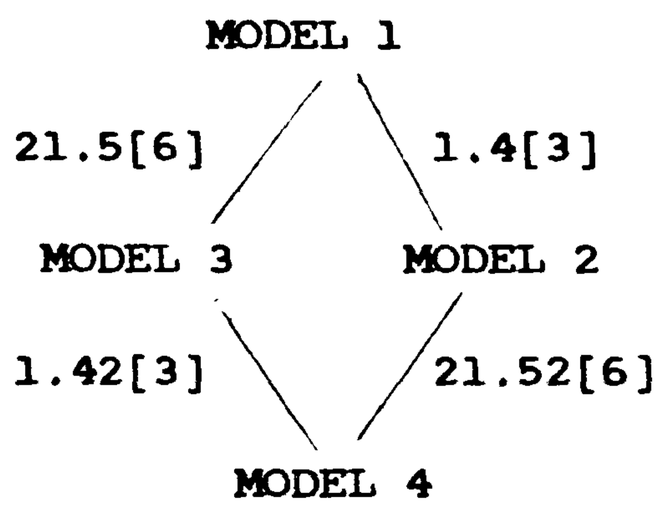
In this diagram we have (parameter)/[integer] where (parameter) ≡ chi-square due to parameter not common to models (m) and (m+1). [Integer] ≡ degrees of freedom due to this uncommon term.

We get;

RULE 2(TRUE POSITIVES)



RULE 3(TRUE POSITIVES)



Of particular interest the chi-squared value for $(\alpha\beta)_{ij}$ is very small, smaller than to be expected even if all the output media were the same. We can think of $\sum_k X_{ijk}$ (sum over observers for given Decision and output medium) as representing effect of output media on given Decision.

The small value of $\chi^2(\alpha\beta)$ could be due to $\sum_k X_{1jk} \approx k_1$ and $\sum_k X_{2jk} \approx k_2$; k_1, k_2 are constants. (2.3.1)

Here $\chi^2(\alpha\beta) \approx \phi$ means accepting model 1, which says that

the Decision is completely independent of observer and output media. Looking back at the contingency table, for the extreme case where $\chi^2(\alpha\beta) = 0$, we could have $X_{1jk} \approx c_1$ and $X_{2jk} \approx c_2$ for all j,k (and c_1, c_2 are constants). This could explain (2.3.1) above. If this is true, then $\mathrm{Var}(\sum_k X_{1jk})$ and $\mathrm{Var}(\sum_k X_{2jk})$ could both be smaller than expected. Likewise Variance of each X_{ijk} (for given i,j) could be smaller than expected. One possible explanation is that each X_{ijk} is not just a simple Binomial but instead is a mixture of 5 binomials. Therefore instead of

$$X_{ijk} \sim \mathrm{Bin}(50, \mu_{ijk})$$

we might have

$$X^{*}_{ijk} \sim \sum_{l=1}^{5} Y_l$$

where $Y_l \sim \mathrm{Bin}(10, \delta_l)$ and Y_l are independent variables then

$$\mathrm{Var}(X_{ijk}) = 50\,\mu_{ijk}(1-\mu_{ijk}) > 10\sum_{l=1}^{5}\delta_l(1-\delta_l) = \mathrm{Var}(X^{*}_{ijk})$$

if $\mu_{ijk} = \sum_{l=1}^{5}\delta_l/5$. This is because the function

$f(\theta) = \theta(1-\theta)$ is concave where we have $f[\frac{1}{n}(\theta_1+\ldots+\theta_n)] \geq \frac{1}{n}[f(\theta_1) +\ldots+ f(\theta_n)]$.

This mixture of binomials could arise from the possibility that for the true Positives an observer looking at any output medium was in fact looking for cold spots of five different visibilities or INTENSITIES.

It seems sensible that an observers decision would depend on how clearly he could see the cold spot (that is the level of visibility of the cold spot). This could directly effect P(T) where we recall that,

P(T) = Prob(Decide cold spot present|POSITIVE CASES) and hence could explain the very small chi-squared value of $\alpha\beta$ for

RULE 2 and RULE 3.

(B) Additional Information

Detailed information concerning intensity of cold spots was fortunately available, which we obtained from Eadie et al (1980). This 'more detailed' data set includes the intensity of the cold spot for every combination of response B_l, observer and output medium. As shown in Section (1.1.A), there are five levels of intensity depending on the length of time an 'absorber' has been applied (see Eadie et al (1980)). In fact, each level of intensity correspond to ten responses (i.e. ten B_l).

To investigate the effect of intensity we first construct tables of intensity against responses (B_l) for every observer and output medium combination, see table [2(III)]. [We regard intensity as zero for false positives simply because there are no cold spots]. All 28 observer/output medium tables seem to exhibit similar trends. Clearly the observers more frequently 'see' the cold spots as intensity increases. They in fact become very sure of their decisions for P100.0 (strongest intensity). We propose to exclude P100.0 from future analysis since its inclusion could mask the variability of observers response for P0.0, P12.5,...,P75.0. (See Table 2(iii) for definitions).

Rather than include intensity as a fourth variable in our log-linear model, we instead consider the logit model

$$L_{IJK} = \log\left[\frac{\Delta_{IJK}}{1-\Delta_{IJK}}\right] \qquad (2.3.2)$$

where Δ_{IJK} = probability observer K decide cold spot present given intensity I and output medium J. We note, for e.g. Everitt (1977), Chapter 5, that the logit model is equivalent to a log-linear model with higher order terms, therefore the ideas of section (2.1) and (2.2) carry over into the present investigation. Advantages in using binary response logit models are discussed in Cox (1970).

P0.0	P12.5	P25.0	P50.0	P75.0	P100.0		
0	0	0	8	8	10	B1	
2	0	0	1	1	0	B2	
8	1	3	1	0	0	B3	AS/OM1
24	4	6	0	1	0	B4	
16	5	1	0	0	C	B5	
0	0	0	1	4	10	B1	
0	0	0	5	4	0	B2	
2	1	0	2	1	0	B3	PH/OM1
30	2	2	2	1	0	B4	
18	7	8	0	0	0	B5	
0	0	0	5	6	10	B1	
0	0	0	3	3	0	B2	
9	0	4	2	0	0	B3	TH/OM1
25	2	4	0	1	0	B4	
16	8	2	0	0	0	B5	
0	0	0	5	6	10	B1	
0	0	0	4	2	0	B2	
3	0	1	1	2	0	B3	FA/OM1
25	1	2	0	0	0	B4	
22	9	7	0	0	0	B5	
0	0	0	8	6	10	B1	
0	0	0	0	3	0	B2	
1	0	0	0	0	0	B3	GS/OM1
4	0	0	2	0	0	B4	
45	10	10	0	1	0	B5	
0	0	0	6	7	10	B1	
3	0	0	2	1	0	B2	
23	1	3	0	0	0	B3	EL/OM1
18	5	5	2	2	0	B4	
6	4	2	0	0	0	B5	
0	0	1	7	6	10	B1	
0	0	0	1	2	0	B2	
12	1	3	0	1	0	B3	MW/OM1
30	6	0	2	0	0	B4	
8	3	6	0	1	0	B5	

TABLE [2(iii)] Observer response over INTENSITY.

Note(1): Observers are AS, PH, TH, FA, GS, EL, MW

Note(2): OM1 = output medium one (DIGITAL X-RAY)

Note(3): P0.0 = zero intensity (false positives)

P12.5, ..., P100.0 are increasing levels

of intensity for true positives

Note(4): B1, .., B5 are responses (see section (1.1.A))

P0.0	P12.5	P25.0	P50.0	P75.0	P100.0		
0	0	0	3	9	10	B1	
1	0	2	4	0	0	B2	
3	2	3	1	0	0	B3	AS/OM2
19	5	4	2	0	0	B4	
27	3	1	0	1	0	B5	
0	0	0	0	0	9	B1	
0	0	0	1	8	1	B2	
0	0	1	6	1	0	B3	PH/OM2
25	6	6	3	1	0	B4	
25	4	3	0	0	0	B5	
0	0	0	3	8	10	B1	
0	0	0	5	1	0	B2	
7	0	5	2	0	0	B3	TH/OM2
20	8	5	0	1	0	B4	
23	2	0	0	0	0	B5	
0	0	0	0	7	10	B1	
0	0	0	5	1	0	B2	
1	0	3	2	1	0	B3	FA/OM2
20	3	4	2	1	0	B4	
29	7	3	1	0	0	B5	
0	0	1	5	8	10	B1	
1	0	1	2	0	0	B2	
1	0	0	0	1	0	B3	GS/OM2
2	0	1	0	0	0	B4	
46	10	7	3	1	0	B5	
0	0	1	6	9	10	B1	
8	4	2	0	1	0	B2	
17	2	4	2	0	0	B3	EL/OM2
20	4	2	1	0	0	B4	
5	0	1	1	0	0	B5	
0	0	0	2	7	10	B1	
0	0	1	2	1	0	B2	
3	0	1	3	0	0	B3	MW/OM2
21	2	4	2	2	0	B4	
26	8	4	1	0	0	B5	

TABLE [2(iii)] continued

Note(5): OM2 = output medium two (ANALOGUE X-RAY)

P0.0	P12.5	P25.0	P50.0	P75.0	P100.0		
0	0	0	4	8	10	B1	
0	0	1	3	1	0	B2	
11	3	4	2	0	0	B3	AS/OM3
32	7	4	1	1	0	B4	
7	0	1	0	0	0	B5	
0	0	0	0	0	10	B1	
0	0	0	1	7	0	B2	
2	0	1	6	3	0	B3	PH/OM3
27	5	7	3	0	0	B4	
21	1	2	0	0	0	B5	
0	0	0	4	8	10	B1	
0	0	0	6	2	0	B2	
15	3	3	0	0	0	B3	TH/OM3
24	3	4	0	0	0	B4	
11	4	3	0	0	0	B5	
0	0	0	0	5	10	B1	
0	0	0	4	4	0	B2	
1	0	1	3	0	0	B3	FA/OM3
27	4	5	3	1	0	B4	
22	6	4	0	0	0	B5	
1	0	0	7	8	10	B1	
2	0	0	0	1	0	B2	
0	0	0	0	0	0	B3	GS/OM3
9	1	0	1	0	0	B4	
38	9	10	2	1	0	B5	
0	0	0	5	9	10	B1	
5	1	2	3	0	0	B2	
16	0	3	0	1	0	B3	EL/OM3
21	6	4	2	0	0	B4	
8	3	1	0	0	0	B5	
0	0	0	5	5	10	B1	
2	0	2	4	2	0	B2	
5	3	1	1	1	0	B3	MW/OM3
29	5	6	0	2	0	B4	
14	2	1	0	0	0	B5	

TABLE [2(iii)] continued

Note(6): OM3 = output medium three (POLAROID)

P0.0	P12.5	P25.0	P50.0	P75.0	P100.0		
0	0	0	7	8	10	B1	
0	0	2	1	1	0	B2	
6	2	5	0	1	0	B3	AS/OM4
22	5	3	2	0	0	B4	
22	3	0	0	0	0	B5	
0	0	0	1	3	9	B1	
0	1	0	4	5	1	B2	
4	1	3	3	2	0	B3	PH/OM4
32	7	6	2	0	0	B4	
14	1	1	0	0	0	B5	
0	0	0	3	6	10	B1	
0	0	0	4	1	0	B2	
4	1	4	1	3	0	B3	TH/OM4
17	3	4	2	0	0	B4	
29	6	2	0	0	0	B5	
0	0	0	1	3	10	B1	
0	0	0	5	6	0	B2	
0	0	3	4	0	0	B3	FA/OM4
17	3	3	0	0	0	B4	
33	7	4	0	1	0	B5	
0	0	0	6	6	10	B1	
0	0	0	2	3	0	B2	
0	0	0	0	0	0	B3	GS/OM4
0	0	0	0	0	0	B4	
50	10	10	2	1	0	B5	
0	0	0	5	7	10	B1	
1	2	1	3	1	0	B2	
8	0	6	0	1	0	B3	EL/OM4
22	5	3	2	0	0	B4	
19	3	0	0	1	0	B5	
0	0	0	7	6	10	B1	
2	0	2	0	1	0	B2	
8	1	2	1	1	0	B3	MW/OM4
29	5	5	2	1	0	B4	
11	4	1	0	1	0	B5	

TABLE [2(iii)] continued

Note(7): OM4 = output medium four (VIDICAM)

MODEL	Rule 3 Deviance (Df)	Rule 4 Deviance (Df)
1A. $L_{IJK} = u + \alpha_I + \beta_J + \gamma_K$ (#)	203.8 (126)	228.1 (126)
2A. $L_{IJK} = u + \alpha_I + \beta_J + \gamma_K + (\alpha\beta)_{IJ}$	175.6 (114)	173.0 (114)
3A. $L_{IJK} = u + \alpha_I + \beta_J + \gamma_K + (\alpha\gamma)_{IK}$	138.2 (102)	188.2 (102)
4A. $L_{IJK} = u + \alpha_I + \beta_J + \gamma_K + (\beta\gamma)_{JK}$	159.6 (108)	165.2 (108)
5A. $L_{IJK} = u + \alpha_I + \beta_J + \gamma_K + (\alpha\beta)_{IJ} + (\alpha\gamma)_{IK}$	109.2 (90)	130.2 (90)
6A. L_{IJK} = all terms except $(\alpha\beta\gamma)_{IJK}$	66.33 (72)	60.48 (72)

TABLE [2(iv)]: LOGIT MODELS

(1) I denote category I of intensity

J " " J of output medium

K " " K of observer

(2) Df = 'degrees of freedom'

(3) (#) For an interpretation of the terms, see Appendix (2.1)

RULE THREE								
10.0	2.0	9.0	3.0	1.0	26.0	12.0	OM1	P0.0
4.0	0.0	7.0	1.0	2.0	25.0	3.0	OM2	
11.0	2.0	15.0	1.0	3.0	21.0	7.0	OM3	
6.0	4.0	4.0	0.0	0.0	9.0	10.0	OM4	
1.0	1.0	0.0	0.0	0.0	1.0	1.0	OM1	P12.5
2.0	0.0	0.0	0.0	0.0	6.0	0.0	OM2	
3.0	0.0	3.0	0.0	0.0	1.0	3.0	OM3	
2.0	2.0	1.0	0.0	0.0	2.0	1.0	OM4	
3.0	0.0	4.0	1.0	0.0	3.0	4.0	OM1	P25.0
5.0	1.0	5.0	3.0	2.0	7.0	2.0	OM2	
5.0	1.0	3.0	1.0	0.0	5.0	3.0	OM3	
7.0	3.0	4.0	3.0	0.0	7.0	4.0	OM4	
10.0	8.0	10.0	10.0	8.0	8.0	8.0	OM1	P50.0
8.0	7.0	10.0	7.0	7.0	8.0	7.0	OM2	
9.0	7.0	10.0	7.0	7.0	8.0	10.0	OM3	
8.0	8.0	8.0	10.0	8.0	8.0	8.0	OM4	
9.0	9.0	9.0	10.0	9.0	8.0	9.0	OM1	P75.0
9.0	9.0	9.0	9.0	9.0	10.0	8.0	OM2	
9.0	10.0	10.0	9.0	9.0	10.0	8.0	OM3	
10.0	10.0	10.0	9.0	9.0	9.0	8.0	OM4	
RULE FOUR								
34.0	32.0	34.0	28.0	5.0	44.0	42.0	OM1	P0.0
23.0	25.0	27.0	21.0	4.0	45.0	24.0	OM2	
43.0	29.0	39.0	28.0	12.0	42.0	36.0	OM3	
28.0	36.0	21.0	17.0	0.0	31.0	39.0	OM4	
5.0	3.0	2.0	1.0	0.0	6.0	7.0	OM1	P12.5
7.0	6.0	8.0	3.0	0.0	10.0	2.0	OM2	
10.0	9.0	6.0	4.0	1.0	7.0	8.0	OM3	
7.0	9.0	4.0	3.0	0.0	7.0	6.0	OM4	
9.0	2.0	8.0	3.0	0.0	8.0	4.0	OM1	P25.0
9.0	7.0	10.0	7.0	3.0	9.0	6.0	OM2	
9.0	8.0	7.0	6.0	0.0	9.0	9.0	OM3	
10.0	9.0	8.0	6.0	0.0	10.0	9.0	OM4	
10.0	10.0	10.0	10.0	10.0	10.0	10.0	OM1	P50.0
10.0	10.0	10.0	9.0	7.0	9.0	9.0	OM2	
10.0	10.0	10.0	10.0	8.0	10.0	10.0	OM3	
10.0	10.0	10.0	10.0	8.0	10.0	10.0	OM4	
10.0	10.0	10.0	10.0	9.0	10.0	9.0	OM1	P75.0
9.0	10.0	10.0	10.0	9.0	10.0	10.0	OM2	
10.0	10.0	10.0	10.0	9.0	10.0	10.0	OM3	
10.0	10.0	10.0	9.0	9.0	9.0	9.0	OM4	

TABLE 2(v) Counts of decide cold spot present.

Note(1): Columns (from the left) are observers AS, PH, TH, ..., MW respectively.

Note(2): OM1 = DIGITAL X-RAY, OM2 = ANALOGUE X-RAY, OM3 = POLAROID, OM4 = VIDICAM.

Note(3): P0.0 = FALSE POSITIVES. P12.5, ..., P75.0 are increasing levels of intensities for true positives.

The logit models considered are given in Table 2(iv). For an interpretation of the terms in the models, see Appendix (2.1).

In Table 2(v) is the counts of 'decide cold spot present' given RULE 3 and RULE 4. We omit rule two as there are several rows and columns of zeros.

(C) Results

For RULE 3 model (5A) was accepted with a Deviance of 109.2 for 90 degrees of freedom. For RULE 4, model (6A) was accepted with a deviance of 60.48 for 72 degrees of freedom. We could therefore say that there exists 'some' structure in our data. However using different RULES lead to different models hence different interpretations. We also see this from Table [2(v)]; intensity four and five for RULE 4 give almost always correct decisions when compared to RULE 3. It could be sensible to remove intensity four and five for RULE 4 when detecting differences in output media. We will however concentrate mainly on model (5A) using RULE 3.

Accepting model (5A) for RULE 3 means that we can 'compare' the four output media in discriminating a given positive intensity from zero intensity, through the $(\alpha\beta)_{IJ}$ terms; and this 'comparison' is illustrated in Appendix (2.1).

In GLIM, the constraints on the parameters are such that any term involving the first category of each variable is given the value zero. For e.g.: $(\alpha\beta)_{I1}$, $(\alpha\gamma)_{1K}$ and $(\alpha\gamma)_{11}$ are all zero, and we shall call such terms DEFAULT ZEROS.

If a particular parameter estimate and its standard error is large, this will indicate some 'unusual' feature of the data. A possible occurrence of large parameter estimates is when a whole row of the contingency table (i.e. data) consists of only zeros.

Without loss of generality, denote the five categories of intensity as I_1, I_2,...,I_5. Similarly, J_1,...,J_4 and K_1,...,K_7 for the four output media and seven observers respectively. We use this notation in table [2(vi)].

The first row, third column of Table [2(vi)] is read as follows.

$$\hat{(\alpha\beta)}_{24} > \hat{(\alpha\beta)}_{22} > \hat{(\alpha\beta)}_{23} > \text{ZERO} \qquad (2.3.3)$$

where '$\wedge$' denote estimates. In other words, given that we compare intensity 2 (i.e. weakest intensity for true positives) with intensity 1 (i.e. the false positives which is a DEFAULT ZERO), output medium four (i.e. Vidicam) is better [see Appendix (2.1) for meaning of better] than output medium one (in this case Digital X-ray). Similarly Analogue X-ray and Polaroid are better than Digital X-ray. Further, from (2.3.3) Vidicam (relative to Digital) is better than Analogue (relative to Digital). Finally, the symbol '*' used as a superscript for J_4 (row one, column three of Table [2(vi)]) means $\hat{(\alpha\beta)}_{24}$ is significantly different from zero. This in turn means that Vidicam is significantly different from digital.

In (2.3.3) all comparisons for output media is with respect to Digital X-ray. To compare, say Vidicam and Polaroid, we do so by redefining the default zero. Essentially this means that we rearrange Table [2(v)] for RULE 3 such that for a given intensity the first row will correspond to output medium three. The results are given in Table [2(vi)].

Let the symbol >> mean significantly different e.g. "Vidicam >> Digital" means Vidicam significantly different from Digital. The results of Table [2(vi)] gives;

I_2: Vidicam >> Digital

I_3: Vidicam >> Digital

Analogue >> Digital

Analogue >> Polaroid

Vidicam >> Polaroid

I_4: No significant difference

I_5: No significant differencc

(2.3.4)

Clearly the effect of intensity is important.

In model (5A), rule 3, we also have the $(\alpha\gamma)_{IK}$ interaction term. We note that $\widehat{(\alpha\gamma)}_{24}$ and $\widehat{(\alpha\gamma)}_{25}$ and their respective standard errors are very large when compared to other estimates of $(\alpha\gamma)_{2K}$. Looking at Table [2(v)] for rule 3; given intensity 2, we have two columns of zeros for observers 4 and 5. This may suggest omitting observers 4 and 5 and refitting the same Logit model. However, we should consider; "is obtaining zero success important for intensity two?". We recall that intensity two corresponds to cold spots that should be the most difficult to detect. What we require then is some 'cost' criteria associated with such decisions which unfortunately we do not have.

Observer 6 showed significant interaction with intensity. Looking down column six, Table [2(v)], rule 3, observer six seems to be "guessing". For the False positives he is wrong for nearly half of his decisions. For intensity two, he seems to be able to detect the cold spot more frequently than others. For intensity 5 his performance could be worse, e.g. when compared to observer 3.

We will not attempt to refit our logit model (5A) with observer 6 removed from the contingency table. This is because we are primarily interested in differences between output media.

Parameter	FIX	Order over second Variable	DEFAULT ZEROS	RULE
$(\alpha\beta)_{IJ}$	I_2	$J_4^* > J_2 > J_3 > J_1$	$I_1 \equiv$ False positive $J_1 \equiv$ Digital X-ray	3
	I_3	$J_4^* > J_2^* > J_3 > J_1$		
	I_4	$J_4 > J_1 > J_2 > J_3$		
	I_5	$J_4 > J_2 > J_3 > J_1$		
	I_2	$J_4 > J_2 > J_3 > J_1$	$I_1 \equiv$ False positive $J_2 \equiv$ Analogue	3
	I_3	$J_4 > J_2 > J_3^* > J_1^*$		
	I_4	$J_4 > J_1 > J_2 > J_3$		
	I_5	$J_4 > J_2 > J_3 > J_1$		
	I_2	$J_4 > J_2 > J_3 > J_1$	$I_1 \equiv$ False positive $J_3 \equiv$ Polaroid	3
	I_3	$J_4^* > J_2^* > J_3 > J_1$		
	I_4	$J_4 > J_1 > J_2 > J_3$		
	I_5	$J_4 > J_2 > J_3 > J_1$		

Table [2(vi)]: Comparing $(\alpha\beta)_{IJ}$ terms.

Note: The symbol '*' means the corresponding $(\widehat{\alpha\beta})_{IJ}$ is significantly different from zero, i.e. $\{(\widehat{\alpha\beta})_{IJ} \pm 2\ \widehat{S.E.}\ (\widehat{\alpha\beta})_{IJ}\}$ excludes zero. Details for Rule 3 only are given in Table 2(vii).

Note: $J_1 \equiv$ Digital X-ray
$J_2 \equiv$ Analogue X-ray
$J_3 \equiv$ Polaroid
$J_4 \equiv$ Vidicam

$I_1 \equiv$ False positives
$I_2, \ldots, I_5$ are increasing intensities for true positives.

Parameter $(\alpha\beta)_{3J}$	Estimate	Standard Error of estimate	Default zeros
J ≡ Analogue	1.379	0.4747	
J ≡ Polaroid	0.3413	0.4778	Digital
J ≡ Vidicam	1.902	0.4804	
J ≡ Digital	-1.379	0.4747	
J ≡ Polaroid	-1.037	0.4647	Analogue
J ≡ Vidicam	0.5231	0.4643	
J ≡ Digital	-0.3413	0.4778	
J ≡ Analogue	1.037	0.4647	Polaroid
J ≡ Vidicam	1.560	0.4703	

Table [2(vii)] GLIM's parameter estimates and their standard errors for Intensity Three (RULE 3). The default zero for intensity is False positives.

CHAPTER 3:

SUMMARY AND FURTHER DISCUSSIONS

In section (1.3.A) we used the criterion.

$$D_{IJ} = [P(T) - P(F)]_I - [P(T) - P(F)]_J \qquad (3.1.1)$$

where, for example I ≡ Vidicam, J ≡ Polaroid. In section (2.3.C), In particular Appendix (2.1), we used the criterion,

$$T_{IJ} = W_I - W_J \qquad (3.1.2)$$

where $W_I = \log\left[\frac{P(T)}{1-P(T)}\right]_I - \log\left[\frac{P(F)}{1-P(F)}\right]_I$

and W_J being similarly defined. As in (3.1.1) we can have I ≡ Vidicam and J ≡ Polaroid say.

Consider the point $\underline{z}$ = (1,0) ≡ (P(T),P(F)) in the plane defined by P(T) and P(F). Using criterion (3.1.1) we geometrically look for the line,

$$P(T) - P(F) = c$$

corresponding to a given output medium that is closest to $\underline{z}$. Using criterion (3.1.2), we seek the curve W_I (for given I) that is closest to the point $\underline{z}$ (see figure (A) in Appendix 2.1). The methods of comparing output media is therefore in a way, geometrically, similar.

The criteria (3.1.1) and (3.1.2) depends on the RULE used (see Table [1(I)]). We have not been able to establish the uniqueness of these RULES. Nevertheless, in Chapter 1, we did obtain some significant comparisons of the output media for RULE 3 and RULE 4 (see Table 1(iv)). RULE 4 in Table [1(iv)] indicate,

Vidicam >> Digital with a frequency 3/7

Vidicam >> Polaroid " " 3/7

Analogue >> Digital " " 4/7

Analogue >> Polaroid " " 2/7

Polaroid >> Digital " " 2/7

where '>>' denote "significantly better than"; and a frequency 3/7 means "three out of seven observers gave significant interval estimates". Let SIE denote "significant interval estimate" for a particular pair of output media. Further, let T = "number of observers who get a SIE" in Table [1(iv)]. Therefore

$$T \sim Bin(7, 0.05)$$

Thus $P(T \geq 2) = 0.044$

Since P(SIE) = 0.05, clearly the frequencies 2/7, 3/7, 4/7 are significant with respect to observers.

We should really do a multiple comparison of the $\hat{D}_{IJ}$ terms in Table [1(iv)]. However, for RULE 4, since most of the negative $\hat{D}_{IJ}$ have upper limits close to zero, we do not expect significant results using any multiple comparison techniques.

Having obtained significant results for RULE 4 in Chapter 1, we considered only RULE 3 in Chapter 2, since the RULE used is not unique. We used logit models in Chapter 2 and the relevant results are given in Table [2(vi)]. The significant comparisons are given in (2.3.4).

If we can ignore the criterion of significant interval estimates, in the relevant methods of Chapters 1 and 2, we seem to have the following.

Vidicam > Analogue > Polaroid > Digital, where > denote "better than".

In terms of the construction of the R.O.C. curve (section 1.1.B), using different rules (Table 1(I)) meant that each point (P(T), P(F)) will have a different interpretation of the 'variables' effecting the consequent observers response. In Chapter 2 this meant different logit models for different rules. Clearly we should have a more definite idea of 'costs' associated with making decisions corresponding to different rules.

The analysis of Chapter 2 showed that the response "decide cold spot present" is dependent on the observer himself, the output medium used and the intensity. It is still reasonable to consider a fourth variable that MIGHT effect observer's response, viz: the individual film or X-ray film itself. Perhaps some "inconsistency" in processing the film could lead to some 'variablity' between individual films. However, the prospect of fitting a logit model to a four-way array of binomial probabilities with only 1's and zero's as the cell entries is clearly not promising.

APPENDIX (1.1)*

ESTIMATED PROBABILITY (TRUE POSITIVES)

AS	PH	TH	FA	GS	EL	MW		
0.52	0.30	0.42	0.42	0.48	0.46	0.48	B1	
0.04	0.18	0.12	0.12	0.06	0.06	0.06	B2	
0.10	0.08	0.12	0.08	0.00	0.08	0.10	B3	DIGITAL
0.22	0.14	0.14	0.06	0.04	0.28	0.16	B4	X-RAY
0.12	0.30	0.20	0.32	0.42	0.12	0.20	B5	
0.44	0.18	0.42	0.34	0.48	0.52	0.38	B1	
0.12	0.20	0.12	0.12	0.06	0.14	0.08	B2	
0.12	0.16	0.14	0.12	0.02	0.16	0.08	B3	ANALOGUE
0.22	0.32	0.28	0.20	0.02	0.14	0.20	B4	X-RAY
0.10	0.14	0.04	0.22	0.42	0.04	0.26	B5	
0.44	0.20	0.44	0.30	0.50	0.48	0.40	B1	
0.10	0.16	0.16	0.16	0.02	0.12	0.16	B2	
0.18	0.20	0.12	0.08	0.00	0.08	0.12	B3	POLAROID
0.26	0.38	0.14	0.26	0.04	0.24	0.26	B4	
0.02	0.06	0.14	0.20	0.44	0.08	0.06	B5	
0.50	0.26	0.38	0.28	0.44	0.44	0.46	B1	
0.08	0.22	0.10	0.22	0.10	0.14	0.06	B2	
0.16	0.18	0.18	0.14	0.00	0.14	0.10	B3	VIDICAM
0.20	0.30	0.18	0.12	0.00	0.20	0.26	B4	
0.06	0.04	0.16	0.24	0.46	0.08	0.12	B5	

ESTIMATED PROBABILITY (FALSE POSITIVES)

AS	PH	TH	FA	GS	EL	MW		
0.00	0.00	0.00	0.00	0.00	0.00	0.00	B1	
0.04	0.00	0.00	0.00	0.00	0.06	0.00	B2	
0.16	0.04	0.18	0.06	0.02	0.46	0.24	B3	DIGITAL
0.48	0.60	0.50	0.50	0.08	0.36	0.60	B4	X-RAY
0.32	0.36	0.32	0.44	0.90	0.12	0.16	B5	
0.00	0.00	0.00	0.00	0.00	0.00	0.00	B1	
0.02	0.00	0.00	0.00	0.02	0.16	0.00	B2	
0.06	0.00	0.14	0.02	0.02	0.34	0.06	B3	ANALOGUE
0.38	0.50	0.40	0.40	0.04	0.40	0.42	B4	X-RAY
0.54	0.50	0.46	0.58	0.92	0.10	0.52	B5	
0.00	0.00	0.00	0.00	0.02	0.00	0.00	B1	
0.00	0.00	0.00	0.00	0.04	0.10	0.04	B2	
0.22	0.04	0.30	0.02	0.00	0.32	0.10	B3	POLAROID
0.64	0.54	0.48	0.54	0.18	0.42	0.58	B4	
0.14	0.42	0.22	0.44	0.76	0.16	0.28	B5	
0.00	0.00	0.00	0.00	0.00	0.00	0.00	B1	
0.00	0.00	0.00	0.00	0.00	0.02	0.04	B2	
0.12	0.08	0.08	0.00	0.00	0.16	0.16	B3	VIDICAM
0.44	0.64	0.34	0.34	0.00	0.44	0.58	B4	
0.44	0.28	0.58	0.66	1.00	0.38	0.22	B5	

(*) Note: For convenience denote observers as;

{AS, PH, TH, ..., MW} ≡ {1, 2, ..., 7}

Appendix (1.2)

To investigate any ordering of θ_i, ϕ_i (i=1,...,5)

We construct a Table of $\hat{\theta}_i/\hat{\phi}_i$ for all observer/output combinations. These ratios are given in Table (A) from which there is some indication of $\hat{\theta}_i/\hat{\phi}_i$ decreasing from B_1 to B_5. Let us ignore the ratios in brackets (i.e. 0/0) in Table (A). We now ask, for example VIDICAM and MW, is the value of 0.45 'significantly' different from 0.55? This is equivalent to asking is $(\theta_4/\phi_4) < (\theta_5/\phi_5)$: or does an interval for

$$\log\left[\frac{\theta_4/\theta_5}{\phi_4/\phi_5}\right] \text{ contain zero?}$$

We use an asymptotic result (Bishop et al 1975, Theorem 14.6.4)

$$\sum_{i=1}^{k} c_i \log x_i \;\dot{:}\; \text{Normal}\left[\Sigma c_i \log \theta_i,\ \Sigma\left[\frac{c_i^2}{m_i}\right]\right] \qquad (*)$$

where $\Sigma c_i \log \theta_i$ is log contrast probabilities from a multinomial table with,

observed counts x_i (i=1,...,k)

expected counts m_i

and probabilities θ_i

and where

$$\sum_{i=1}^{k} c_i = 0,\ m_i = N\theta_i$$

We will estimate m_i in (*) by x_i itself.

As an example, we have the interval estimate for

$$\log\left[\frac{\theta_4}{\theta_5}\right] - \log\left[\frac{\phi_4}{\phi_5}\right] \text{ as}$$

$$\log x_4 - \log x_5$$

$$-(\log y_4 - \log y_5) \pm 1.96 \left[\frac{1}{x_4} + \frac{1}{x_5} + \frac{1}{y_4} + \frac{1}{y_5}\right]^{1/2}$$

where x_i = observed counts in category B_i for P(T)

	AS	PH	TH	FA	GS	EL	MW
	∞	∞	∞	∞	∞	∞	∞
	1.00	∞	∞	∞	∞	1.00	∞
DIGITAL	0.63	2.0	0.67	1.33	0.0	0.17	0.42
X-RAY	0.46	0.23	0.28	0.12	0.5	0.78	0.27
	0.38	0.83	0.63	0.73	0.47	1.00	1.25
	∞	∞	∞	∞	∞	∞	∞
	6.00	∞	∞	∞	3.00	0.88	∞
ANALOGUE	2.00	∞	1.00	6.00	1.00	0.47	1.33
X-RAY	0.58	0.64	0.70	0.50	0.50	0.35	0.48
	0.19	0.28	0.09	0.38	0.46	0.40	0.50
	∞	∞	∞	∞	25.00	∞	∞
	∞	∞	∞	∞	0.50	1.20	4.00
POLAROID	0.82	5.00	0.40	4.00	(0/0)	0.25	1.20
	0.41	0.70	0.29	0.48	0.22	0.57	0.45
	0.14	0.14	0.64	0.45	0.58	0.50	0.21
	∞	∞	∞	∞	∞	∞	∞
	∞	∞	∞	∞	∞	7.00	1.50
VIDICAM	1.33	2.25	2.25	∞	(0/0)	0.88	0.63
	0.45	0.47	0.53	0.35	(0/0)	0.45	0.45
	0.14	0.14	0.28	0.36	0.46	0.21	0.55

Table (A)

Values of $\hat{\theta}_i/\hat{\phi}_i$, i=1,...,5

Note (1): Entries that are underlined show increasing order of $\hat{\theta}_i/\hat{\phi}_i$.

	OBSERVER	I	Interval estimate for $\log\{\theta_I/\Phi_I)/(\theta_{I+1}/\Phi_{I+1})\}$
DIGITAL X-RAY	PH	4	-2.34, -0.20
	TH	4	-1.95, 0.35
	FA	4	-3.16, -0.44
	GS	3	involves log (zero)
	EL	3	-2.77, -0.23
	EL	4	-1.58, 1.08
	MW	4	-2.76, -0.33
ANALOGUE X-RAY	EL	4	-1.99, 1.72
	MW	4	-1.05, 0.96
POLAROID	TH	4	-2.05, 0.49
	GS	4	-2.58, 0.66
	EL	3	-2.13, 0.48
VIDICAM	FA	4	-1.17, 1.11
	MW	4	-1.39, 0.99

TABLE (B)

where y_l = observed counts in category B_l for P(F)

The interval estimates for log $[(\theta_l/\phi_l)/(\theta_{l+1}/\phi_{l+1})]$ corresponding to an increasing order of $\hat{\theta}_i/\hat{\phi}_i$ is given in Table (B). We of course want zero to be in these intervals. This is the case for Analogue X-ray, Polaroid and Vidicam. However, for DIGITAL X-ray four of the intervals do not contain zero, suggesting that there may not be a simple ordering of the $\hat{\theta}_i/\hat{\phi}_i$ for this output medium.

Appendix (1.3)

$$\text{Let } \lambda(\underline{x}) = \frac{P(\underline{X} = \underline{x}|\theta_1)}{P(\underline{X} = \underline{x}|\theta_2)} = \frac{P_1(\underline{X} = \underline{x})}{P_2(\underline{X} = \underline{x})}$$

where P(.) denotes a probability function.

Define a Likelihood ratio (LR) rule as

Make, decision (1) if $\lambda(\underline{x}) > C$
decision (2) if $\lambda(\underline{x}) < C$
(1) or (2) if $\lambda(\underline{x}) = C$
→ for given constant C

Let $\alpha = P_1$ (LR decides (2))

$\beta = P_2$ (LR decides (1))

Let another rule (perhaps not an LR) have corresponding errors α^* and β^*.

Theorem: $\alpha + c\beta \leq \alpha^* + c\beta^*$

Proof: Divide $\underline{x}$ - space into 4 disjoint sets.

LR decides (1)	A	B
LR decides (2)	D	G

other rule decides

(1) (2)

where for example the subset A is when both rules make

decision (1).

Using the usual set notation 'ϵ' and 'U' to mean 'an element of' and 'union' respectively.

when $\underline{x} \epsilon B \Rightarrow \lambda(\underline{x}) \geq C$

$\underline{x} \epsilon D \Rightarrow \lambda(\underline{x}) \leq C$.

By definition,

$\alpha = P_1(D \cup G)$

$\alpha^* = P_1(B \cup G)$

$\beta = P_2(A \cup B)$

$\beta^* = P_2(A \cup D)$

and since A, B, D, G are disjoint we have for example,

$P_1(D \cup G) = P(D) + P(G)$

Thus $\alpha^* - \alpha = P_1(B) - P_1(D)$

$\beta^* - \beta = P_2(D) - P_2(B)$

Thus $\alpha^* + C\beta^* - (\alpha + C\beta) = P_1(B) - CP_2(B)$

$- [P_1(D) - CP_2(D)]$

$= \sum_{\underline{x} \epsilon B} \{P_1(\underline{X} = \underline{x}) - C\, P_2(\underline{X} = \underline{x})\}$

$- \sum_{\underline{x} \epsilon D} \{P_1(\underline{X} = \underline{x}) - C\, P_2(\underline{X} = \underline{x})\}$

≥ 0 since $\lambda(\underline{x}) \geq C$ when $\underline{x} \epsilon B$

$\lambda(\underline{x}) \leq C$ when $\underline{x} \epsilon D$

Appendix (1.4)

We illustrate the derivation of the interval estimate (for given observer)

$$\hat{D}_{IJ} \pm 1.96 \sqrt{\hat{Var}(\hat{D}_{IJ})}$$

where $D_{IJ} = [P(T) - P(F)]_I - [P(T) - P(F)]_J$ (*)

$= D_I - D_J$

where I $\equiv$ Vidicam and J $\equiv$ Polaroid, say.

As an example we consider RULE 2, and for the moment obtain

the estimated variance of.

$$\hat{D} = \hat{P}(T) - \hat{P}(F) = (\hat{\theta}_1 + \hat{\theta}_2) - (\hat{\phi}_1 + \hat{\phi}_2)$$

$$\text{Thus Var } (\hat{D}) = \text{Var } \left[\frac{m_1}{N} + \frac{m_2}{N} - \frac{n_1}{N} - \frac{n_2}{N}\right]$$

since $\hat{\theta}_i = \frac{m_i}{N}$, $\hat{\phi}_i = \frac{n_i}{N}$ (see Table 1(ii) and Table 1(iii))

and where N = 50.

Now we assume m_i independent of n_i. $\forall_i$

$$\text{Thus } \text{Var}(\hat{D}) = \frac{1}{N^2}\{\text{Var}(m_1+m_2) + \text{Var}(n_1+n_2)\}$$

$$= \frac{1}{N}\{(\theta_1+\theta_2)(1-\theta_1-\theta_2)+(\phi_1+\phi_2)(1-\phi_1-\phi_2)\}$$

The estimate $\hat{\text{Var}}(\hat{D})$ is then obtained by estimating θ_i and ϕ_i with $\frac{m_i}{N}$ and $\frac{n_i}{N}$ respectively.

We calculate the relevant $\hat{\text{Var}}(\hat{D})$ for Polarold and Vidicam and substitute into (*) to get $\hat{D}_{IJ}$. We next make another assumption: "An observer's rule (Table 1(I)) is independent of the output media." Thus, $\text{Var}(\hat{D}_{IJ}) = \text{Var}(\hat{D}_I - \hat{D}_J)$

$$= \text{Var}(\hat{D}_I) + \text{Var}(\hat{D}_J)$$

$$\text{and } \hat{\text{Var}}(\hat{D}_{IJ}) = \hat{\text{Var}}(\hat{D}_I) + \hat{\text{Var}}(\hat{D}_J)$$

Appendix (2.1)

(A) Interpretation of terms in logit-model

Let U_{IJK} be the logit of true probability of deciding cold spot present, for intensity I, output medium J and observer K.

Consider,

$$U_{IJK} = \mu + \alpha_I + \beta_J + \gamma_K + (\alpha\beta)_{IJ} + (\alpha\gamma)_{IK} + (\beta\gamma)_{JK} \qquad (*)$$

From (*) we see that,

$$\begin{aligned} U_{IJ1} - U_{1J1} - U_{I11} + U_{111} \\ = U_{IJ2} - U_{1J2} - U_{I12} + U_{112} \\ . \\ . \\ . \\ . \\ . \\ = U_{IJK} - U_{1JK} - U_{I1K} + U_{11K} \end{aligned} \qquad (**)$$

Therefore, given $(\alpha\beta\gamma)_{IJK} = 0$ for all I, J, K, the quantity

$$U_{IJK} - U_{1JK} - U_{I1K} + U_{11K} \qquad (***)$$

is independent of K (i.e. observer) for all I, J. We can arbitrarily choose the value of K such that the quantity given in (***) is fixed. In particular, GLIM uses the definition,

$$(\alpha\beta)_{IJ} = (U_{IJ1} - U_{1J1}) - (U_{I11} - U_{111}) \qquad (****)$$

We note that intensity one was defined to be the FALSE POSITIVES (see section (1.1.B)). So, (****) is a measure of difference between output medium J and output medium one, when discriminating intensity I from intensity one; which does not depend on observer.

We note that model (5A), in table [2(iv)] has $(\beta\gamma)_{JK}$ equal to zero. This does not alter the meaning of the $(\alpha\beta)_{IJ}$ term given in (****).

A similar discussion is given in McCullagh and Nelder (1983),

section (3.5.2), where the analogy between using the 'usual' constraints (e.g. $\Sigma\alpha_I = 0$) and GLIM's parameterisation (e.g. $\alpha_1 = 0$) is shown.

GLIM also use the definition,

$$\beta_J = U_{1J1} - U_{111}$$

a quantity we are not interested in as it is not a contrast involving two intensities.

(B) Geometry in the (P(T), P(F)) plane

$$\text{Let } \exp(U_{1J1}) = \frac{y}{1-y} \text{ and } \exp(U_{1J1}) = \frac{x}{1-x}$$

where U_{IJK} is defined in (*).

Consider the plot of,

$$\log\left[\frac{y}{1-y}\right] - \log\left[\frac{x}{1-x}\right] = c \qquad (+)$$

where c is a constant. The graph (+) is equivalent to,

$$y = \frac{dx}{1 + (d-1)x} \qquad (++)$$

where $d = \exp(c)$.

To illustrate, we plot the graph (++) which is given in figure (A) for various values of d. Clearly, as d increases (i.e. c increases) the graph "moves" towards the left and upper boundaries of the unit-square (i.e. towards the lines x = 0 and y = 1).

But $y \equiv P(T)$ and $x \equiv P(F)$, see section (1.1.B) for definitions. From figure (A) the graph associated with larger values of P(T) (given P(F)) have larger values of c. In particular, let $(P(T), P(F))_V$ correspond to a point on the graph with $c = c_V$ for Vidicam say, and $(P(T), P(F))_A$ is the point for Analogue X-ray where $c=c_A$, say. If c_V minus c_A is positive we regard Vidicam as better than Analogue X-ray for a given intensity I.

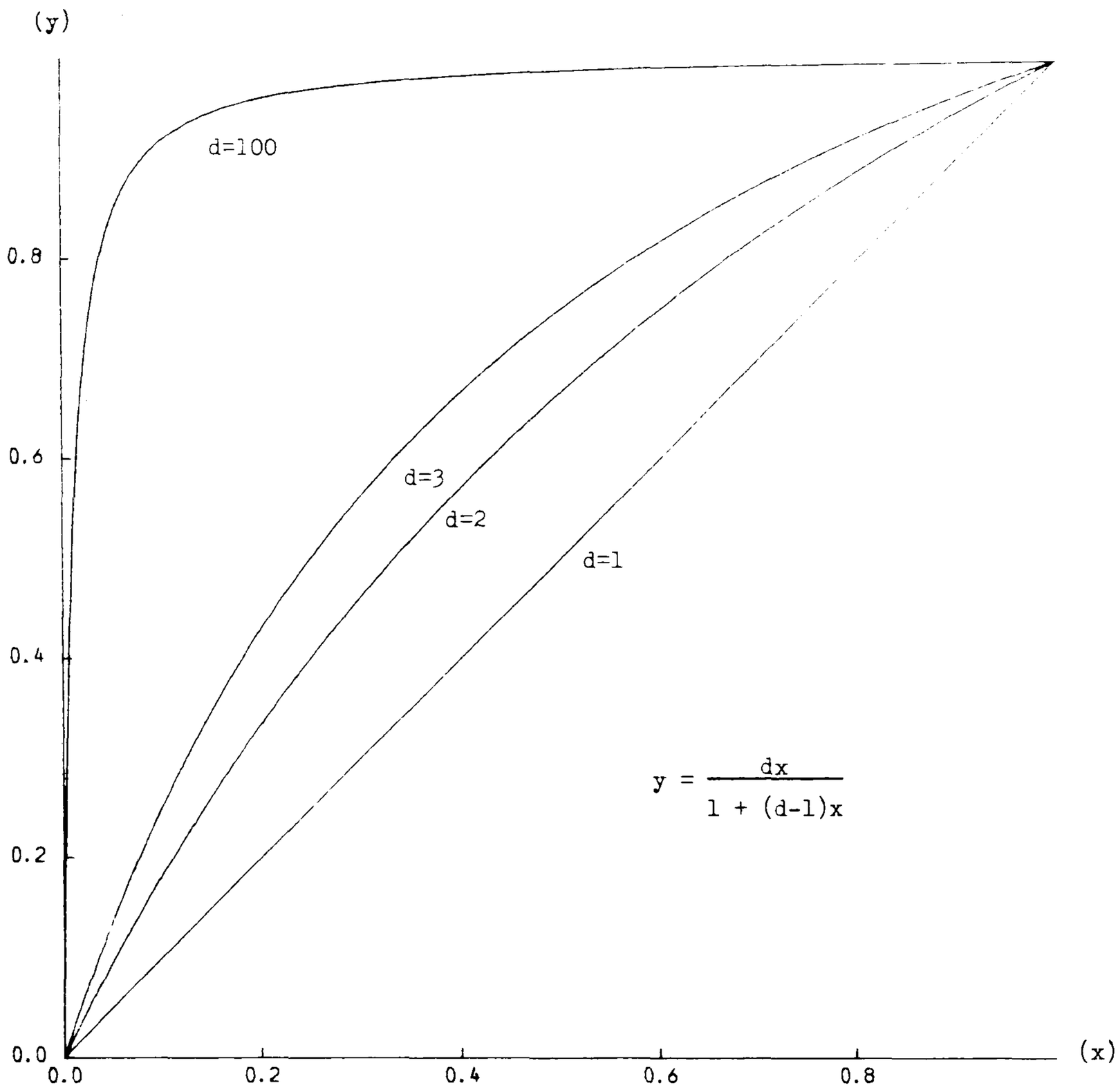


FIGURE (A)

Appendix (2.2)

We rewrite equation (2.1.3).

$$P(X_{ijk}) = P(X_{.jk})P(X_{1jk}|x_{.jk}) \qquad (*)$$

where $X_{ijk} \sim P_o(e_{ijk})$.

Clearly, $X_{.jk} \sim P_o(e_{.jk})$ where the 'dot' in the subscripts indicate summation over the subscript i.

Consider the model,

$$\log(e_{ijk}) = u + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + (\alpha\beta)_{ij}$$

$$= \alpha_i + \eta_{jk} + (\alpha\beta)_{ij}$$

where $\eta_{jk} = u + \beta_j + \gamma_k + (\beta\gamma)_{jk}$.

Since $P(X_{1jk}|x_{.jk}) \sim \text{Bin}(x_{.jk}, \phi_{jk})$

$$\text{Thus } P(X_{1jk}|x_{.jk}) = \prod_{j,k} \begin{bmatrix} x_{.jk} \\ x_{1jk} \end{bmatrix} \phi_{jk}^{x_{1jk}} (1-\phi_{jk})^{x_{2jk}}$$

where $\phi_{jk} = e_{1jk}/(e_{1jk} + e_{2jk})$.

Here $\prod_{jk}$ denote the product over all (j,k). As illustrated in section (2.1) the term η_{jk} cancels out in ϕ_{jk}.

Further, $P(X_{.jk})$ is the product (over j,k) of

$$\frac{\exp[-e_{.jk}]\, e_{.jk}^{x_{.jk}}}{x_{.jk}!}$$

which depends only on $e_{.jk}$ and

$$e_{.jk} = \exp[\alpha_1 + (\alpha\beta)_{1j} + \eta_{jk}]$$

$$+ \exp[\alpha_2 + (\alpha\beta)_{2j} + \eta_{jk}]$$

When we estimate $\{\alpha_i, (\alpha\beta)_{ij}\}$ we also have the estimate of $P(X_{1jk}|x_{.jk})$. We can use these estimates of $(\alpha_i, (\alpha\beta)_{ij})$ to estimate $P(X_{.jk})$ since $\hat{\eta}_{jk}$ need only satisfy $\hat{e}_{.jk} = 50$, and can be arbitrarily estimated. In other words, $P(X_{1jk}|x_{.jk})$ can be estimated independently of $P(X_{.jk})$. Therefore, the ratio of two Poisson likelihoods, i.e. $P(X_{ijk})$ is equivalent to the ratio of two binomials, i.e. $P(X_{1jk}|x_{.jk})$

Chapter 1: TWO POPULATION DISCRIMINATION

(1.1) Introduction

Let $Np(\underline{\mu}_i, \Omega_i)$ denote the p-variate normal distribution with mean $\underline{\mu}_i$ and covariance matrix Ω_i. Further, let Π_i denote population i and $f(\underline{x}|\Pi_i)$ be the probability density function of $\underline{x}$ given Π_i.

In the two population discrimination problem we consider a particular case.

$$f(\underline{x}|\Pi_i) \sim Np(\underline{\mu}_i, \Omega_i) \text{ for } i=1,2.$$

The classical discriminant analysis largely concentrates on the odds ratio $f(\Pi_1|\underline{x})/f(\Pi_2|\underline{x})$.

We have,

$$\frac{f(\Pi_1|\underline{x})}{f(\Pi_2|\underline{x})} = \frac{f(\Pi_1)}{f(\Pi_2)} \cdot \frac{f(\underline{x}|\Pi_1)}{f(\underline{x}|\Pi_2)}$$

where $f(\Pi_i)$ is the prior probability of sampling an observation $\underline{x}$ from Π_i. It is assumed that the $f(\Pi_i)$'s are known or could be estimated from suitable data.

The parameter of interest is;

$$\theta(\underline{x}) = \log_e \left[\frac{f(\underline{x}|\Pi_1)}{f(\underline{x}|\Pi_2)}\right]$$

which will be referred to as the 'log-odds'.

(1.2) Estimation of $\theta(\underline{x})$

We seek inference for $\theta(\underline{x})$. From the 'training' samples of sizes n_1 and n_2 we can calculate $\underline{\bar{x}}_i$ and $S_i (i=1,2)$ where $\underline{\bar{x}}_i$ is the sample mean and S_i the corrected sum of squares and cross-products matrix, for Π_i.

One approach in estimating $\theta(\underline{x})$ is to 'plug-in' $\underline{\bar{x}}_i$ and S_i (assuming $\Omega_1 \neq \Omega_2$) into the formula for $\theta(\underline{x})$ by letting $\hat{\underline{\mu}}_i = \underline{\bar{x}}_i$ and $\hat{\Omega}_i = kS_i$, where k = appropriate constant.

The plug-in method has been critisized by Aitchison, Habbema and Kay (1977) where they suggested using a Bayesian approach

based on predictive distributions. Aitchison, Habbema and Kay (1977) argue that the 'plug-in' method yields extreme estimates of the odds, possibly the consequence of not taking into account the repeated sampling properties of $\hat{\theta}(\underline{x})$.

Moran and Murphy (1979) showed that the study in Aitchison, Habbema and Kay (1977) ignored the bias involved in estimating the odds when using the 'plug-in' method. Having made adjustments for bias, Moran and Murphy (1979) showed that the 'plug-in' method is then more comparable to the methods used by Aitchison, Habbema and Kay (1977).

Estimates of the odds ratio using the three methods discussed in this section, for a particular discrimination problem, can vary considerably. Even for a particular method we do not expect the corresponding odds to be estimated well under all situations.

(1.3) Assessing Discriminant Rules

A frequently used criterion in assessing a discriminant rule is the unconditional probability of misclassification, see for example Lachenbruch (1975). The use of unconditional probabilities is normally associated with situations where a decision has to be made one way or the other, and is not necessarily informative about the uncertainty involved in a particular decision.

Critchley and Ford (1985) consider possible cases as illustrated in figures (1(i)) and (1(ii)). In figure (1(i)), even with low misclassification probabilities the point A will involve a decision made with great uncertainty. In contrast, in figure (1(ii)), a decision for point B will be made with great certainty despite the high misclassification probabilities involved. Clearly it is important to consider conditionally, on $\underline{x}$, the uncertainty with which any decision is made. This uncertainty will involve $\theta(\underline{x})$ itself and the extra uncertainty associated with the estimation of $\theta(\underline{x})$. Critchley and

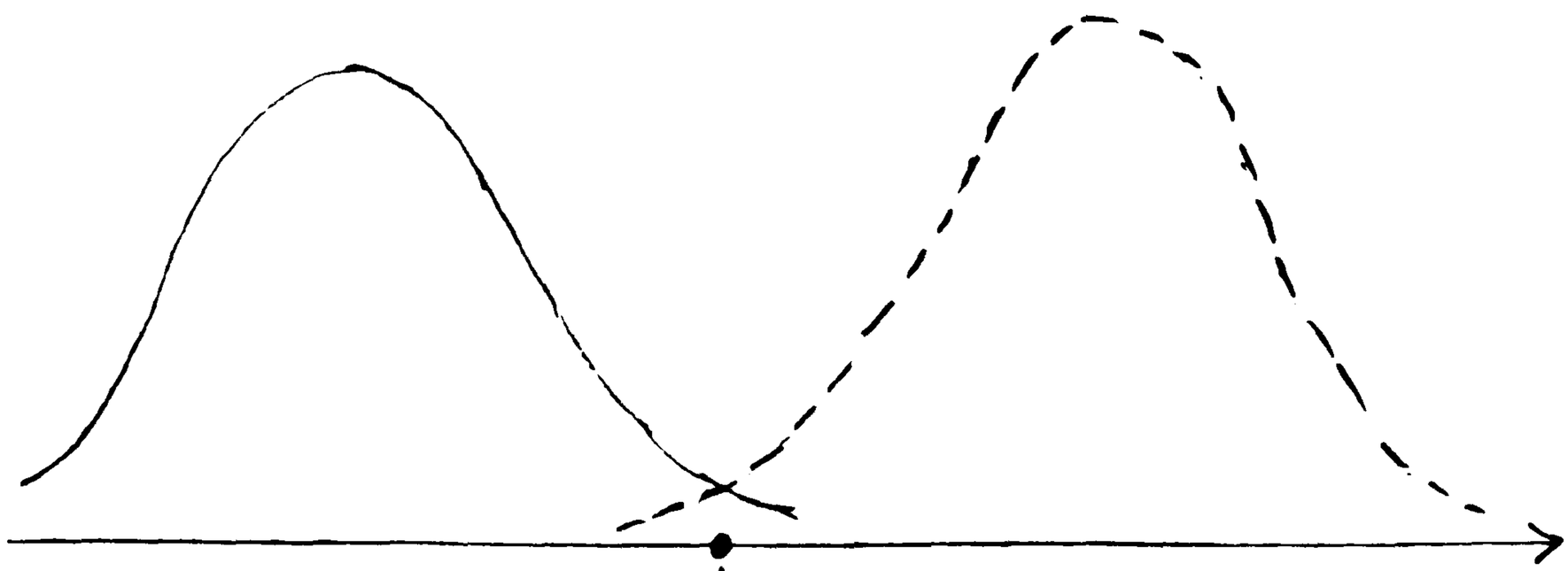


Figure (1(i)): Some decisions made with great uncertainty.

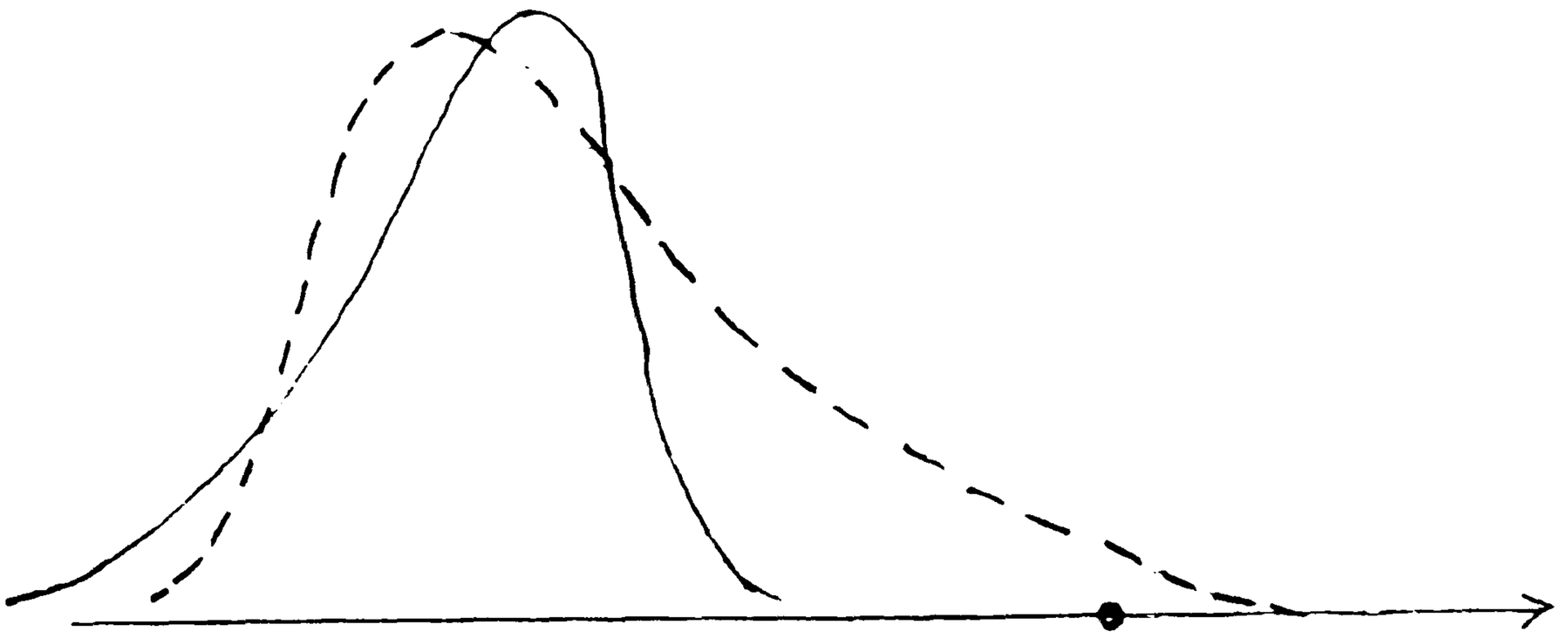


Figure (1(ii)): Some decisions made with great certainty.

Ford (1985) studied the latter type of uncertainty.

(1.4) The Equal Covariance Case

Using unbiased estimates of $e(\underline{x})$ from Moran and Murphy (1979), the problem of interval estimation for $e(\underline{x})$ was considered by Critchley and Ford (1985) for the multivariate normal case with $\Omega_1=\Omega_2$. An exact variance for the unbiased estimate $\hat{e}(\underline{x})$ is given in Critchley and Ford (1984).

Critchley and Ford (1985) show that useful information concerning the p-dimensional discrimination problem can be displayed geometrically in a two-dimensional plot. This informative plot displays, in particular,

(i) the discriminant scores

(ii) $\mathrm{Var}(\hat{e}(\underline{x}))$, representing uncertainty in decisions for $\underline{x}$-points associated with a given score.

Critchley and Ford (1985) studied two approximate interval estimates for $e(\underline{x})$, both of which are based on large sample properties.

(1.5) The unequal covariance case

The aim of this thesis is to extend the work done in Critchley and Ford (1985) for the particular case of unequal covariance matrices.

When the covariance matrices are not equal, we have the problem of the increase in the number of parameters involved in the two population discrimination problem. As a result, we study the use of approximate interval estimation methods for $e(\underline{x})$.

Another approach to this problem from a Bayesian viewpoint, is discussed in Rigby (1982). The results in this thesis are based on non-Bayesian methods.

CHAPTER 2

DERIVATION OF SOME USEFUL RESULTS

(2.1) Introduction and Notation

Let Π_i denote population i, and let $f(\underline{x}|\Pi_i)$ be the probability density function of $\underline{x}$ in population i. Thus,

$$f(\underline{x}|\Pi_i) \sim Np\ (\underline{\mu}_i,\ \Omega_i)\quad (i=1,2)$$

where $Np(\underline{\mu}_i,\ \Omega_i)$ denotes the p-variate normal distribution with mean $\underline{\mu}_i$ and covariance matrix Ω_i.

We seek inference for the log-odds θ, viz:-

$$\theta(\underline{x}) = \theta(\underline{\mu}_1,\ \underline{\mu}_2,\ \Omega_1,\ \Omega_2|\underline{x}) = \log_e\left[\frac{f(\underline{x}|\Pi_1)}{f(\underline{x}|\Pi_2)}\right]$$

In this chapter we gather together derivations of some technical results which will be used in later chapters. We will use the letters E and NE to denote $\Omega_1 = \Omega_2$ and $\Omega_1 \neq \Omega_2$ respectively. Without loss of generality, denote the covariance matrix for Π_i as

(i) Ω when $\Omega_1 = \Omega_2$ and

(ii) Ω_i when $\Omega_1 \neq \Omega_2$

Further define,

$$\left.\begin{aligned}
&h_i^2(\underline{x}) = (\underline{x} - \underline{\mu}_i)^T\Omega^{-1}(\underline{x} - \underline{\mu}_i) \quad (i=1,2)\\
&\alpha_i^2(\underline{x}) = (\underline{x} - \underline{\mu}_i)^T\Omega_i^{-1}(\underline{x} - \underline{\mu}_i) \quad (i=1,2)\\
&\Delta^2 = (\underline{\mu}_1 - \underline{\mu}_2)^T\Omega^{-1}(\underline{\mu}_1 - \underline{\mu}_2)\\
&\Phi(\underline{x}) = \frac{1}{2}\left[h_1^2(\underline{x}) + h_2^2(\underline{x})\right]\\
&\underline{\delta}^T = (\Delta,\ 0,\ldots,0)^T
\end{aligned}\right\} \qquad (2.1.1)$$

We can think of $h_i^2(\underline{x})$ as a measure of 'atypicality', and $\Phi(\underline{x})$ as 'average atypicality'.

Estimates will be marked with a "hat-sign" e.g. the estimate of $\alpha_{i(}^2\underline{x})$ is $\hat{\alpha}_i^2(\underline{x})$.

Clearly,

$$\theta_E(\underline{x}) = \frac{1}{2}\left[h_2^2(\underline{x}) - h_1^2(\underline{x})\right]$$

and

$$\theta_{NE}(\underline{x}) = \frac{1}{2}(\alpha_2^2(\underline{x}) - \alpha_1^2(\underline{x})) - \frac{1}{2}\log_e \{|\Omega_1|/|\Omega_2|\}$$

where $|\Omega_i|$ denotes the determinant of Ω_i.

As was mentioned in Chapter 1, unbiased estimates of $\theta_E(x)$ and $\theta_{NE}(\underline{x})$ can be obtained from Moran and Murphy (1979). The unbiased estimates are:

$$\hat{\theta}_E(\underline{x}) = \frac{1}{2}p\left[\frac{1}{n_1} - \frac{1}{n_2}\right] +$$

$$+ (n_1 + n_2 - p-3)\,(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)^T S^{-1} \left[\underline{x} - \frac{1}{2}(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)\right] \qquad (2.1.2)$$

where $\bar{\underline{x}}_i$ is the mean of the i^{th} 'training' sample,

S is the pooled corrected sum of squares and cross products matrix (SSP),

n_i is the sample size (i=1,2).

Since the data is normally distributed;

$$\bar{\underline{x}}_i \sim Np(\underline{\mu}_i, \frac{1}{n_i}\Omega) \text{ and } S \sim Wp(n_1 + n_2 - 2, \Omega) \qquad (2.1.3)$$

where the latter distribution is the p-variate Wishart distribution.

$$\text{Also, } \hat{\theta}_{NE}(\underline{x}) = -\frac{1}{2}\{\hat{\alpha}_1^2(\underline{x}) - \hat{\alpha}_2^2(\underline{x})\} + \frac{1}{2}p\left[\frac{1}{n_1} - \frac{1}{n_2}\right]$$

$$-\frac{1}{2}\ln[|S_1|/|S_2|]$$

$$+\frac{1}{2}\sum_{j=1}^{p}\left[\psi\left[\frac{n_1-j}{2}\right] - \psi\left[\frac{n_2-j}{2}\right]\right] \qquad (2.1.4)$$

$$\text{where } \bar{\underline{x}}_i \sim Np(\underline{\mu}_i, \frac{1}{n_i}\Omega_i) \quad (i=1,2)$$

$$S_i \sim Wp(n_i - 1, \Omega_i) \quad (i=1,2)$$

$$\hat{\alpha}_i^2(\underline{x}) = (\underline{x} - \bar{\underline{x}}_i)^T(a_i S_i^{-1})\,(\underline{x} - \bar{\underline{x}}_i) \quad (i=1,2) \qquad (2.1.5)$$

$$a_i = n_i - p - 2$$

$\psi(.)$ = digamma function. (Abramowitz and Stegun (1972))

Note the use of S as the pooled corrected SSP matrix when $\Omega_1 = \Omega_2$. and S_i as the corrected SSP matrix for Π_i when $\Omega_1 \neq \Omega_2$.

To calculate the digamma function we make use of the algorithm from Bernardo (1976).

It is of interest to calculate the variances of $\hat{e}_E(\underline{x})$ and $\hat{e}_{NE}(\underline{x})$ so that we can express the uncertainty in our point estimators. In Section 2.2 we reproduce the variance of $\hat{e}_E(\underline{x})$, a result from Critchley and Ford (1984). In Section 2.3 we derive an approximate variance for $\hat{e}_{NE}(\underline{x})$.

As there will be a bias in using $\hat{e}_E(\underline{x})$ when we actually have $\Omega_1 \neq \Omega_2$. we derived an approximation for this bias in Section (2.5).

In Section (2.4) we derive an approximation to

$$E((\log|S_i^{-1}|)S_i^{-1})$$

and use this result to obtain another approximate variance for $\hat{e}_{NE}(\underline{x})$. The approximate variances obtained in this chapter will be compared empirically in a later chapter.

(2.2) Variance of $\hat{e}_E(\underline{x})$

The variance of $\hat{e}_E(\underline{x})$ (see Critchley and Ford (1984)) is

$$(N-p)(N-p-3)V(\hat{e}_E(\underline{x})) = (N-p+1)\{e(\underline{x})-\frac{1}{2}(N-1)\left[\frac{1}{n_1}-\frac{1}{n_2}\right]\}^2$$

$$+ (N-p-1)[\phi(\underline{x})\{(N-1)\left[\frac{1}{n_1}+\frac{1}{n_2}\right] + \Delta^2\} - \frac{1}{4}\Delta^4]$$

$$+ \frac{1}{4}(N-1)(N-p-1)\{2p\left[\frac{1}{n_1^2}+\frac{1}{n_2^2}\right] - (N+1)\left[\frac{1}{n_1}-\frac{1}{n_2}\right]^2\}$$

(2.2.1)

We of course need to estimate $V\{\hat{e}_E(\underline{x})\}$. The simulation results of Critchley and Ford (1985) make use of an unbiased estimator of $V(\hat{e}_E)$ viz:-

$$\hat{V}_1(\hat{e}_E) = \frac{\left[(N-p)(N-p-3)\hat{V}\#(\hat{e}_E) - 2\hat{e}_E^2 - 2(N-1)\left[\frac{1}{n_2}-\frac{1}{n_1}\right]\hat{e}_E - f''\right]}{[(N-p)(N-p-3) + (N-p+1)]}$$

(2.2.2)

where $f''' = \dfrac{f - \left[\dfrac{(N-1)(N-p-1)p}{n_1 n_2}\right]}{(N-p)}$

$$f = \frac{(N-1)(N-p-1)p}{2} \left[\frac{1}{n_1^2} + \frac{1}{n_2^2}\right]$$

$$+ p \left[\frac{1}{n_2} - \frac{1}{n_1}\right]^2 \frac{(N-1)}{2}$$

$$N = n_1 + n_2 - 2$$

$\hat{V}\#(\hat{\theta}_E)$ is the "plug-in" estimator (using the minimum variance unbiased estimators of Δ^2, θ_E, $||\underline{x}^*||^2 = \phi(\underline{x}) - \frac{1'}{4}\Delta^2$) of $V(\hat{\theta}_E(\underline{x}))$.

In the next chapter, we carry out a simulation study on the distribution of $\hat{\theta}_E(\underline{x})$, and interval estimation for $\theta(x)$. We note the possibility of $\hat{V}_1(\hat{\theta}_E)$ becoming negative. We will record any such instances, if they occur, and replace $V_1(\hat{\theta}_E)$ by $V_2(\hat{\theta}_E)$ obtained by 'plugging-in' the unbiased estimates of Ω, $\underline{\mu}_1$ and $\underline{\mu}_2$.

(2.3) Variance of $\hat{\theta}_{NE}(\underline{x})$

An approximate variance for $\hat{\theta}_{NE}(x)$ is,

$$AV\{\hat{\theta}_{NE}(\underline{x})\}$$

$$= \sum_{i=1}^{2} \left[\frac{\{\alpha_i^2(\underline{x})\}^2}{2(n_i-p-4)} + \left\{\frac{1}{n_i} - \frac{(n_i-2)}{(n_i-p-1)(n_i-p-4)}\right\} \alpha_i^2(\underline{x}) + \frac{p(n_i-2)}{2(n_i-p-1)(n_i-p-4)}\right] \qquad (2.3.1)$$

Proof: Define,

$$\text{diag}(A1, A2, A3, A4) = \begin{bmatrix} A1 & 0 & 0 & 0 \\ 0 & A2 & 0 & 0 \\ 0 & 0 & A3 & 0 \\ 0 & 0 & 0 & A4 \end{bmatrix}$$

where the matrix 0 is a matrix of zeros with appropriate dimensions.

Next, define vec(M) as the $p(p+1)/2$ vector whose elements are the upper triangular elements of the pxp matrix M.

Let $\beta^T = [\underline{\mu}_1^T, \underline{\mu}_2^T, \text{vec}(\Omega_1^{-1}), \text{vec}(\Omega_2^{-1})]$ (2.3.2)

From (2.1.5) clearly $E(\bar{\underline{x}}_i) = \underline{\mu}_i$.

From Gupta (1968), $E(a_i S_i^{-1}) = \Omega_i^{-1}$ where $a_i = n_i - p - 2$ and $S_i \sim Wp(n_i - 1, \Omega_i)$.

Let $\hat{\beta}^T = [\bar{x}_1^T, \bar{x}_2^T, \text{vec}(a_1 S_1^{-1}), \text{vec}(a_2 S_2^{-1})]$

Therefore $E(\hat{\beta}) = \beta$

Let $\theta_{NE}(x) = g(\beta)$, i.e. a function of β.

We consider an approximation of $g(\hat{\beta})$,

$$g(\hat{\beta}) = g(\beta) + [g'(\beta)]^T (\hat{\beta} - \beta) + \text{higher order terms}$$

$$\text{where } g'(\beta) = \frac{\delta[g(\hat{\beta})]}{\delta \hat{\beta}} \Bigg|_{\hat{\beta} = \beta}$$

Thus $E[g(\hat{\beta})] = g(\beta)$, to first order terms.

Hence,

$$V[g(\hat{\beta})] \doteq [g'(\beta)]^T \text{ cov } (\hat{\beta}) \, [g'(\beta)] \qquad (2.3.3)$$

where 'V', 'cov' denote variance and covariance respectively,

and cov $(\hat{\beta}) = \text{diag} \left[\frac{1}{n_1} \Omega_1, \frac{1}{n_2} \Omega_2, \text{cov}(\text{vec}(a_1 S_1^{-1})), \right.$

$$\left. \text{cov}(\text{vec}(a_2 S_2^{-1})) \right] ,$$

since $\bar{x}_1$, $\bar{x}_2$, S_1, S_2 are independent sets of variables.

We shall define $\text{cov}(\text{vec}(a_i S_i^{-1}))$ shortly.

To maintain consistency in notation, we denote the right hand side of (2.3.3) by $AV\{g(\hat{\beta})\}$. Thus,

$$AV\{g(\hat{\beta})\} = AV\{\hat{\theta}_{NE}(x)\} = [g'(\beta)]^T \text{ cov}(\hat{\beta})[g'(\beta)] \qquad (2.3.4)$$

Clearly $g'(\beta) = \begin{bmatrix} \underline{g_{11}} \\ \underline{g_{12}} \\ \underline{g_{21}} \\ \underline{g_{22}} \end{bmatrix}$

where $g_{1i} = \frac{\delta g}{\delta \bar{x}_i}$ evaluated at $\bar{x}_i = \mu_i$ $\quad (i=1,2)$

$$g_{2i} = \frac{\delta g}{\delta[\mathrm{vec}(a_i S_i^{-1})]} \quad \text{evaluated at } \bar{\underline{x}}_i = \underline{\mu}_i \text{ and } a_i S_i^{-1} = \Omega_i^{-1} \quad (i=1,2)$$

We therefore want

$$AV\{\hat{\theta}_{NE}(\underline{x})\} = g_{11}^T \left[\frac{1}{n_1}\Omega_1\right] g_{11} + g_{12}^T \left[\frac{1}{n_2}\Omega_2\right] g_{12}$$

$$+ g_{21}^T \mathrm{cov}[\mathrm{vec}(a_1 S_1^{-1})]\, g_{21} + g_{22}^T \mathrm{cov}[\mathrm{vec}(a_2 S_2^{-1})]\, g_{22}$$

$$= z_{11} + z_{12} + z_{21} + z_{22} \qquad (2.3.5)$$

We now establish the relevant components for the right hand side of (2.3.5).

(i) $g_{1i} = (-1)^{i+1} a_i S_i^{-1}(\underline{x}-\bar{\underline{x}}_i)$ [evaluated at $\bar{\underline{x}}_i = \underline{\mu}_i$ $a_i S_i^{-1} = \Omega_i^{-1}$]

$$= (-1)^{i+1}\Omega_i^{-1}(\underline{x}-\underline{\mu}_i).$$

(ii) $2g_{2i} = (-1)^i \mathrm{vec}[2(\underline{x}-\underline{\mu}_i)(\underline{x}-\underline{\mu}_i)^T - D_{\underline{b}\underline{b}^T}]$

$$+ (-1)^{i+1} \mathrm{vec}[2\Omega_i - D_{\Omega_i}].$$

where $D_{\underline{b}\underline{b}^T}$ is the diagonal matrix whose i^{th} diagonal element is equal to the i^{th} diagonal element of the matrix $\underline{b}\,\underline{b}^T$, $\underline{b} = \underline{x} - \underline{\mu}_i$. The diagonal matrix D_{Ω_i} is similarly defined.

(i) and (ii) above are standard results, see for example Graybill [1983], chapter 10.

We require two further results.

(v) $T \sim Wp(N, U) \Rightarrow E(T^{-1}) = \frac{U^{-1}}{N-p-1}$ if $N-p-1>0$

a result from Gupta (1968).

(vi) $T \sim Wp(N, U)$ and let $T^{-1} = \{t^{ij}\}$, $U^{-1} = \{u^{ij}\}$.

From Siskind (1972) we have,

$$E(t^{ij}t^{rs}) = \frac{(N-p-2)u^{ij}u^{rs} + u^{ir}u^{js} + u^{is}u^{jr}}{(N-p)(N-p-1)(N-p-3)}$$

we first use (v) and (vi), and recall that

$$\frac{1}{a}S \sim Wp(n-1, \frac{1}{a}\Omega)$$

for each Π_i

Let $aS^{-1} = (s^{ij})$, $a\Omega^{-1} = (\delta^{ij})$ and $\Omega = \{\omega_{ij}\}$

Thus $\text{cov}(s^{ij}, s^{rs}) = E(s^{ij}s^{rs}) - E(s^{ij})\,E(s^{rs})$

$$= \frac{2\delta^{ij}\delta^{rs} + (t-2)\,[\delta^{ir}\delta^{js} + \delta^{is}\delta^{jr}]}{(t-1)(t-2)^2(t-4)} \qquad (2.3.6)$$

where $t = n-p$.

We next use (2.3.6) to calculate z_{21} and z_{22}. Without loss of generality, for the time being, drop the subscript 'i' from $\underline{\mu}_i$, Ω_i, z_{2i}, t_i, n_i, a_i and $\underline{b}_i$.

Let $\Omega - (\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})^T = \begin{bmatrix} & & [q] \\ & \underline{r} & \\ [q] & & \end{bmatrix}$

where [q] contains upper triangle elements of $\Omega - (\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})^T$; and $\underline{r}$ contains the corresponding diagonal elements.

Let $\text{cov}(\hat{\Omega}^{-1}) = \text{cov}(aS^{-1}) = \left[\begin{array}{c|c|c} B_{11} & B_{11} & B_{12} \\ \hline B_{11} & B_{11} & B_{12} \\ \hline B_{12}^T & B_{12}^T & B_{22} \end{array}\right]$

such that $B_{11} = \text{cov}(s^{ij}, s^{rs})$ where $\frac{\delta q}{\delta s^{ij}} \in [q]$ and $\frac{\delta q}{\delta s^{rs}} \in [q]$

$B_{12} = \text{cov}(s^{ij}, s^{rs})$ " " $\in [q]$ " " $\in \underline{r}$

$B_{22} =$ " " " $\in \underline{r}$ " " $\in \underline{r}$

Let $[\underline{q}]$ denote the elements of [q] expressed as a vector. We want,

$$z_2 = \begin{bmatrix} [\underline{q}] \\ \underline{1} \\ 2 \ \underline{r} \end{bmatrix}^T \begin{bmatrix} B_{11} & B_{12} \\ & \\ B_{12}^T & B_{22} \end{bmatrix} \begin{bmatrix} [\underline{q}] \\ \underline{1} \\ 2 \ \underline{r} \end{bmatrix}$$

$$= \sum_{i\le j} \sum_{r\le s} \sum \sum \frac{\delta q}{\delta [aS^{-1}]}^T [\text{cov}(aS^{-1})] \frac{\delta q}{\delta [aS^{-1}]}$$

$$= [\underline{q}]^T B_{11} [\underline{q}] + [\underline{q}]^T B_{12}\, \underline{r} + \frac{\underline{r}^T B_{22} \underline{r}}{4}$$

And we note that

$$\begin{bmatrix} [g] \\ [g] \\ \underline{r} \end{bmatrix}^T \begin{bmatrix} B_{11} & B_{11} & B_{12} \\ B_{11} & B_{11} & B_{12} \\ B_{12}^T & B_{12}^T & B_{22} \end{bmatrix} \begin{bmatrix} [g] \\ [g] \\ \underline{r} \end{bmatrix} = 4\, z_2$$

Therefore

$$z_2 = \frac{1}{4} \sum_i \sum_j \sum_r \sum_s (\Omega - \underline{b}\,\underline{b}^T)_{ij}(\Omega - \underline{b}\,\underline{b}^T)_{rs}\, \mathrm{cov}(s^{ij}, s^{rs})$$

$$= \frac{1}{4.\mathrm{DENOM}} \sum_i \sum_j \sum_r \sum_s (\omega_{ij} - b_i b_j)(\omega_{rs} - b_r b_s)$$

$$[2\delta^{ij}\delta^{rs} + (t-2)(\delta^{ir}\delta^{js} + \delta^{is}\delta^{jr})]$$

$$= [2pa^2(n-2) + 2a^2(t-1)[\underline{b}^T\Omega^{-1}\underline{b}]^2 - 4a^2(\underline{b}^T\Omega^{-1}\underline{b})(n-2)]/(4.\mathrm{DENOM}) \qquad (2.3.7)$$

see Appendix 2.1 for details.

where $t = n-p$

$a = n-p-2$

$\mathrm{DENOM} = (t-1)(t-2)^2(t-4)$

$\underline{b} = \underline{x} - \underline{\mu}$

We will use the subscript k to denote the various variables from Π_k (k=1,2).

Let $W_k = (\underline{x}-\underline{\mu}_k)^T\Omega_k^{-1}\left[\frac{1}{n_k}\Omega_k\right]\Omega_k^{-1}(\underline{x}-\underline{\mu}_k)$

$$+ \frac{1}{4} \sum_i \sum_j \sum_r \sum_s (\Omega_k - \underline{b}_k\underline{b}_k^T)_{ij}\,(\Omega_k - \underline{b}_k\underline{b}_k^T)_{rs}\, \mathrm{cov}_{\Pi_k}(s^{ij}, s^{rs})$$

Substitute W_k into (2.3.5) and we have

$$\mathrm{AV}(\hat{\theta}_{NE}(\underline{x})) = W_1 + W_2 \qquad (2.3.8)$$

Finally, by substituting the appropriate z_{2i} [from (2.3.7)] into (2.3.8) we have the required form of $\mathrm{AV}(\hat{\theta}_{NE}(\underline{x}))$ that is given in (2.3.1). The estimate of $\mathrm{AV}(\hat{\theta}_{NE}(\underline{x}))$ when we substitute $\underline{\mu}_i$ and Ω_i by $\bar{\underline{x}}_i$ and $\frac{1}{(n_i-p-2)}S_i$ (i=1,2) will be called

$$\hat{\mathrm{AV}}_1(\hat{\theta}_{NE}(\underline{x}))$$

(2.4) Another approximate variance for $\hat{\theta}_{NE}(\underline{x})$ using an approximation to $E((\log|S^{-1}|)S^{-1})$

(A) Introduction

In Section (2.3) we have obtained an asymptotic approximation to $Var(\hat{\theta}_{NE}(\underline{x}))$, using a linear approximation to $\hat{\theta}_{NE}(\underline{x})$. We now pursue a more exact approximation based on the idea:

$$\hat{\theta}_{NE}(\underline{x}) = \text{constant} + \frac{1}{2}[\hat{\alpha}_2^2(\underline{x}) - \hat{\alpha}_1^2(\underline{x}) - \log|S_1| + \log|S_2|]$$

From (2.1.4) and (2.1.5) we have $\bar{\underline{x}}_i \# S_i$ (#denotes independence).

Further, the two samples are independent. Therefore

$$\alpha_1^2(\underline{x}) \# \alpha_2^2(\underline{x}),\ \alpha_1^2(\underline{x}) \# \log|S_2|$$

and $\alpha_2^2(\underline{x}) \# \log|S_1|,\quad \log|S_1| \# \log|S_2|$

Clearly,

$$Var(\hat{\theta}_{NE}(\underline{x})) = \frac{1}{4}\begin{bmatrix} var(\hat{\alpha}_1^2) + var(\hat{\alpha}_2^2) \\ +var(\log|S_1|) + var(\log|S_2|) \\ +2cov(\hat{\alpha}_1^2, \log|S_1|) + 2cov(\hat{\alpha}_2^2, \log|S_2|) \end{bmatrix} \quad (2.4.1)$$

(Here $\hat{\alpha}_1^2 \equiv \hat{\alpha}_1^2(\underline{x})$).

All the terms on the right-hand side of (2.4.1) are known exactly except $\text{cov}(\hat{\alpha}_i^2, \log|S_i|)$. From

$\text{cov}(X, Y) \equiv E(XY) - E(X)E(Y)$, X,Y are random variables,

$Var[\hat{\theta}_{NE}(\underline{x})]$ would be known exactly if and only if we knew $E(\hat{\alpha}_i^2 \log|S_i|)$ exactly. This amounts to knowing the matrix;

$$E((\log|S_i^{-1}|).S_i^{-1}) \text{ exactly.} \quad (2.4.2)$$

Unfortunately we do not have this latter result and therefore propose to find an approximation to (2.4.2).

(B) An approximation to $E((\log|S_i^{-1}|)S_i^{-1})$

Here we drop the subscript i in S_i for convenience. We define the distinct elements of S^{-1} by $[s^{gh} | g \le h]$.

Let $f_{ij}(S^{-1}) \equiv (\log|S^{-1}|)s^{ij}$

Using a second order Taylor's expansion in the elements of S^{-1} about

$E(S^{-1}) \equiv \Gamma^{-1} \equiv \frac{\Omega^{-1}}{(N-p-1)}$ we have,

$$f_{ij}(S^{-1}) \simeq f_{ij}(\Gamma^{-1}) + \sum_{g\leq h} (s^{gh}-\gamma^{gh}) \left[\left[\frac{\delta f_{ij}}{\delta s^{gh}}\right]\Big|_{S^{-1}=\Gamma^{-1}}\right]$$

$$+ \frac{1}{2} \sum_{g\leq h} \sum_{k\leq l} (s^{gh}-\gamma^{gh})(s^{kl}-\gamma^{kl}) \left[\left[\frac{\delta^2 f_{ij}}{\delta s^{gh}\delta s^{kl}}\right]\Big|_{S^{-1}=\Gamma^{-1}}\right]$$

so that

$$E[f_{ij}(S^{-1})] \simeq f_{ij}(\Gamma^{-1}) + 0$$

$$+ \frac{1}{2} \sum_{g\leq h} \sum_{k\leq l} \text{cov}(s^{gh},s^{kl}) \left[\left[\frac{\delta^2 f_{ij}}{\delta s^{gh}\delta s^{kl}}\right]\Big|_{S^{-1}=\Gamma^{-1}}\right]$$

$\text{cov}(s^{gh},s^{kl})$ is given in (2.3.6) and is reproduced here.

$$\text{cov}(s^{gh},s^{kl}) = \frac{2\sigma^{gh}\sigma^{kl} + (N-p-1)(\sigma^{gk}\sigma^{hl} + \sigma^{gl}\sigma^{hk})}{(N-p)(N-p-1)^2(N-p-3)}$$

where $N = n-1$

We now find the second derivative of $f_{ij}(S^{-1})$

$$f_{ij} \equiv (\log|S^{-1}|)s^{ij}$$

$$\text{Thus } \frac{\delta f_{ij}}{\delta s^{gh}} = (2-\delta_{gh})\, s_{gh}\, s^{ij} + \delta_{(gh)(ij)}(\log|S^{-1}|)$$

where δ_{gh} = kronecker delta = (1 if g = h
(0 otherwise

and $\delta_{(gh)(ij)}$ = (1 if (g,h) = (i,j)
(0 if (g,h) $\neq$ (i,j)

$$\frac{\delta^2 f_{ij}}{\delta s^{gh}\delta s^{kl}} = (2-\delta_{gh})\left[\delta_{(kl)(ij)}\right]s_{gh}$$

$$+ (2-\delta_{gh})\, s^{ij}\left(-\frac{1}{2}\right)(2-\delta_{kl})(s_{gk}s_{lh} + s_{gl}s_{kh})$$

$$+ \left[\delta_{(gh)(ij)}\right](2 - \delta_{kl})\, s_{kl}$$

where $\frac{\delta s_{gh}}{\delta s^{kl}}$ is obtained from Result (1), Appendix (2.2)

and $\frac{\delta(\log|S^{-1}|)}{\delta\, s^{kl}}$ is obtained from Result (3), Appendix (2.2)

and the second derivative is to be evaluated at $S^{-1} = \Gamma^{-1}$,

or equivalently at $s_{ij} = (N-p-1)\sigma_{ij}$, $\forall\ i \leq j$.

Thus $E\{f_{ij}(S^{-1})\} \simeq f_{ij}(\Gamma^{-1})$

$$+ \frac{1}{2} \sum_{g \leq h} (2-\delta_{gh})[\sigma_{gh}(N-p-1)]\ \ \mathrm{cov}(s^{gh}, s^{ij})$$

$$+ \frac{1}{2} \sum_{k \leq l} (2-\delta_{kl})\,[(N-p-1)\sigma_{kl}]\ \mathrm{cov}(s^{kl}, s^{ij})$$

$$- \frac{1}{4} \sum_{g \leq h} \sum_{k \leq l} (2-\delta_{gh})(2-\delta_{kl})\,[(N-p-1)\sigma^{ij}(\sigma_{gk}\sigma_{lh} + \sigma_{gl}\sigma_{kh})]$$

$$\mathrm{cov}(s^{gh}, s^{kl}) \qquad (2.4.3)$$

The second and third terms on the right-hand side of (2.4.3) are equal. Their common value is;

$$\frac{1}{2} \sum_{\text{all } g,h} (N-p-1)\sigma_{gh}\ \mathrm{cov}(s^{gh}, s^{ij})$$

$$= (N-1)K\ \sigma^{ij} \qquad (2.4.4)$$

where $K = \{(N-p)(N-p-1)(N-p-3)\}^{-1}$

The last term in the approximation to $E\{f_{ij}(S^{-1})\}$ is;

$$- \frac{1}{4} K\sigma^{ij} \sum_{\text{all } g,h} \sum_{\text{all } k,l} (\sigma_{gk}\sigma_{lh} + \sigma_{gl}\sigma_{kh})$$

$$\{2\sigma^{gh}\sigma^{kl} + (N-p-1)(\sigma^{gk}\sigma^{hl} + \sigma^{gl}\sigma^{hk})\}$$

$$= -\frac{K}{2}\,\sigma^{ij}\,\{p(N-p+1) + p^2(N-p-1)\} \qquad (2.4.5)$$

Collecting (2.4.4) and (2.4.5), putting them back into (2.4.3), we have the desired approximation;

$$E\{(\log|S^{-1}|)S^{-1}\} \simeq (\log|\Gamma^{-1}|)\Gamma^{-1}$$

$$+ \left[\frac{2(N-1)}{(N-p)(N-p-3)}\right] \Gamma^{-1}$$

$$- \left[\frac{p^2(N-p-1)+p(N-p+1)}{2(N-p)(N-p-3)}\right]\Gamma^{-1} \qquad (2.4.6)$$

where $\Gamma^{-1} = \dfrac{\Omega^{-1}}{N-p-1}$

(C) Another approximate Variance for $\hat{\theta}_{NE}(\underline{x})$

We want an approximation to $\mathrm{Var}(\hat{\theta}_{NE}(\underline{x}))$ as given in (2.4.1).

Let $T_i^2 = \frac{n_i(n_i-1)}{a_i} \hat{\alpha}_i^2(\underline{x})$ [See (2.1.5) for definitions]

and $\lambda_i = n_i\alpha_i^2(\underline{x})$

Since $\frac{T_i^2}{(n_i-1)} \frac{(n_i-p)}{p} \sim F(p, n_i-p, \lambda_i)$ [Anderson (1958) Theorem (5.2.2)]

and from Johnson and Kotz (1970b), page 190, we have,

$$\text{Var}[\hat{\alpha}_i^2(\underline{x})] = \frac{2[\alpha_i^2(\underline{x})]^2}{(n_i-p-4)} + \frac{4(n_i-2)\alpha_i^2(\underline{x})}{n_i(n_i-p-4)} + \frac{2p(n_i-2)}{n_i^2(n_i-p-4)} \qquad (2.4.7)$$

From Johnson and Kotz (1970c), page 198,

$$\text{Var}\{\log(|S_i|/|\Omega_i|)\} = \sum_{j=1}^{p} \psi'[\tfrac{1}{2}(n_i-j)] \qquad (2.4.8)$$

where $\psi'(.) \equiv$ trigamma function [Abramowitz and Stegun (1972)].

For our purpose, to calculate $\psi'(.)$, use (for m = integer),

$$\text{(a)}\quad \psi'(m-1) = \frac{\pi^2}{6} - \sum_{k=1}^{m-2} \frac{1}{k^2} \qquad (m=3,4,5,\ldots)$$

[from (6.4.3) and (23.2.24) of Abramowitz and Stegun (1970)].

$$\text{(b)}\quad \psi'(m-\tfrac{1}{2}) = \frac{\pi^2}{2} - 4\sum_{k=1}^{m-1} \frac{1}{(2k-1)^2}$$

[from (6.4.5) of Abramowitz and Stegun (1970)] (2.4.9)

We require the covariance term in (2.4.1). Firstly,

$$\text{cov}[\hat{\alpha}_i^2(\underline{x}),\ \log|S_i|] \equiv E[\hat{\alpha}_i^2(\underline{x}).(\log|S_i|)] - E\{\hat{\alpha}_i^2(\underline{x})\}\ E(\log|S_i|)$$

Again from Johnson and Kotz (1970b) and (1970c),

$$E\{\hat{\alpha}_i^2(\underline{x})\} = \alpha_i^2(\underline{x}) + \frac{p}{n_i} \qquad (2.4.10)$$

$$E\{\log(|S_i|/|\Omega_i|)\} = p\log 2 + \sum_{j=1}^{p} \psi\{\tfrac{1}{2}[n_i-j]\} \qquad (2.4.11)$$

where $\psi(.) \equiv$ digamma function.

Let $B_i \equiv E_{\Omega_i}\{(\log|S_i^{-1}|)S_i^{-1}\}$

we want $E\{\hat{\alpha}_i^2(\underline{x})(\log|S_i|)\}$

$$= E_{\underline{\bar{x}}_i}\{-a_i(\underline{x}-\underline{\bar{x}}_i)^T B_i(\underline{x}-\underline{\bar{x}}_i)\}$$

$$= -a_i\{\underline{x}^T B_i \underline{x} - 2\underline{x}^T B_i \underline{\mu}_i + \underline{\mu}_i^T B_i \underline{\mu}_i + tr\left[B_i\left[\frac{1}{n_i}\Omega_i\right]\right]\}$$

From (2.4.6) we get,

$$B_i \triangleq \frac{t_i}{a_i}\,\Omega_i^{-1}$$

where $t_i = \log(|\Gamma_i^{-1}|) + \frac{2(N_i-1)}{(N_i-p)(N_i-p-3)}$

$$- \left[\frac{p^2(N_i-p-1) + p(N_i-p+1)}{2(N_i-p)(N_i-p-3)}\right] \qquad (2.4.12)$$

where $\Gamma_i^{-1} = \frac{\Omega_i^{-1}}{(N_i-p-1)}$ and $N_i = n_i-1$

Clearly,

$$E\{\hat{\alpha}_i^2(\underline{x})(\log|S_i|)\} \triangleq -t_i[\alpha_i^2(\underline{x}) + \frac{p}{n_i}] \qquad (2.4.13)$$

Combining the results from (2.4.7), (2.4.8), (2.4.10), (2.4.11) and (2.4.13), we now have the sum of variance-covariance terms for population i. Therefore,

$$\text{var}(\hat{\alpha}_i^2(\underline{x})) + \text{var}[\log|S_i|] + 2\,\text{cov}[\hat{\alpha}_i^2(\underline{x}), (\log|S_i|)]$$

$$\triangleq \frac{2[\alpha_i^2(\underline{x})]^2}{(n_i-p-4)} + \left[\frac{4(n_i-2)}{n_i(n_i-p-4)} - 2t_i - 2u_i\right]\alpha_i^2(\underline{x})$$

$$+ \sum_{j=1}^{p}\psi'[\tfrac{1}{2}(n_i-j))] + \frac{2p(n_i-2)}{n_i(n_i-p-4)} - \frac{2pu_i}{n_i} - \frac{2t_i p}{n_i} \qquad (2.4.14)$$

where $u_i = p\log 2 + \sum_{j=1}^{p}\psi[\tfrac{1}{2}(n_i-j)] + (\log|\Omega_i|)$

t_i is given in (2.4.12)

For purposes of notation, let,

$$BV\{\hat{\theta}_{NE}(\underline{x})\}$$

be the approximate variance of $\hat{\theta}_{NE}(\underline{x})$ when we substitute (2.4.14) into (2.4.1) for i=1 and 2.

We will compare empirically the 'performance' of

$AV(\hat{\theta}_{NE}(\underline{x}))$ and $BV(\hat{\theta}_{NE}(\underline{x}))$ in Chapter 3.

(2.5) : Approximation to the bias of $\hat{\theta}_E(\underline{x})$ when $\Omega_1 \neq \Omega_2$

(A) Introduction

We recall that,

$$\theta_{NE}(\underline{x}) = \frac{1}{2}\{\alpha_2^2(\underline{x}) - \alpha_1^2(\underline{x})\} - \frac{1}{2}\log_e\{|\Omega_1|/|\Omega_2|\}$$

If we assumed $\Omega_1 = \Omega_2$, when in fact the covariance matrices are unequal, we would then be using $\hat{\theta}_E(\underline{x})$ to estimate $\theta_{NE}(\underline{x})$.

From (2.1.2),

$$\hat{\theta}_E(\underline{x}) = \frac{1}{2}[\hat{h}_2^2(\underline{x}) - \hat{h}_1^2(\underline{x})]$$

where $\hat{h}_i^2(\underline{x}) = (\underline{x} - \underline{\bar{x}}_i)^T \left[\frac{S_1 + S_2}{n_1 + n_2 - p - 3}\right]^{-1} (\underline{x} - \underline{\bar{x}}_i) - \frac{p}{n_i}$

We want,

$$\theta_{NE}(\underline{x}) - E(\hat{\theta}_E(\underline{x}))$$

$$= \frac{1}{2}\left[\log[|\Omega_2|/|\Omega_1|] + \alpha_2^2(\underline{x}) - \alpha_1^2(\underline{x}) - E(\hat{h}_2^2(\underline{x})) + E(\hat{h}_1^2(\underline{x}))\right]$$

We shall call this BIAS1.

To compute $E(\hat{h}_i^2(\underline{x}))$ we need $E\{(S_1 + S_2)^{-1}\}$. Unfortunately, for $\Omega_1 \neq \Omega_2$ the matrix $(S_1 + S_2)$ does not have a 'simple' distribution. It is therefore difficult to find $E\{(S_1 + S_2)^{-1}\}$ analytically. Instead, we propose to do a Taylor series expansion of $(S_1 + S_2)^{-1}$ to second order in the elements of $(S_1 + S_2)$.

(B) Taylor's expansion

For all $i \leq j$, let $f_{ij}(S) = s^{ij} = (i,j)^{th}$ element of S^{-1}

where S is a symmetric p x p matrix [and $S = S_1 + S_2$].

Let $\Gamma = E(S)$

Clearly $\Gamma = (n_1 - 1)\Omega_1 + (n_2 - 1)\Omega_2$

then; $f_{ij}(S) \simeq f_{ij}(\Gamma) + \sum_{g\leq h} (s_{gh} - \gamma_{gh}) \left[\left[\frac{\delta f_{ij}(S)}{\delta s_{gh}}\right]\right|_{S=\Gamma}\right]$

$$+ \frac{1}{2}\sum_{g\leq h}\sum_{k\leq \ell} (s_{gh} - \gamma_{gh})(s_{k\ell} - \gamma_{k\ell}) \left[\left[\frac{\delta^2 f_{ij}}{\delta s_{gh}\delta s_{k\ell}}\right]\right|_{S=\Gamma}\right]$$

so that,

$E\{f_{ij}(S)\} = E(s^{ij})$

$$\simeq \Gamma^{ij} + 0 + \frac{1}{2}\sum_{g\leq h}\sum_{k\leq \ell} \operatorname{cov}(s_{gh}, s_{k\ell}) \left[\left[\frac{\delta^2 f_{ij}}{\delta s_{gh}\delta s_{k\ell}}\right]\right|_{S=\Gamma}\right] \qquad (2.5.1)$$

Let $s_{ij}^{(k)} = (i,j)^{th}$ element of S_k.

Cov $(s_{gh}, s_{k\ell})$

$= \operatorname{cov}(s_{gh}^{(1)} + s_{gh}^{(2)}, s_{k\ell}^{(1)} + s_{k\ell}^{(2)})$

$= \operatorname{cov}(s_{gh}^{(1)}, s_{k\ell}^{(1)}) + \operatorname{cov}(s_{gh}^{(2)}, s_{k\ell}^{(2)})$ [since S_1 is independent of S_2]

$$= \sum_{\alpha=1}^{2}(n_\alpha - 1)\,[\sigma_{gk}^{(\alpha)}\,\sigma_{h\ell}^{(\alpha)} + \sigma_{g\ell}^{(\alpha)}\,\sigma_{hk}^{(\alpha)}] \text{ (see Appendix (2.2), result 2)} \qquad (2.5.2)$$

where $\sigma_{ij}^{(K)} = (i,j)^{th}$ element of Ω_K

also; $\dfrac{\delta f_{ij}}{\delta s_{gh}} \equiv \dfrac{\delta s^{ij}}{\delta s_{gh}}$

$= -\frac{1}{2}(2 - \delta_{gh})(s^{ig}\,s^{hj} + s^{ih}\,s^{gj})$ (see Appendix 2.2 result 1)

where $\delta_{gh} = \begin{cases} 1 & \text{if } g = h \\ 0 & \text{otherwise.} \end{cases}$

Thus $\dfrac{\delta^2 f_{ij}}{\delta s_{gh}\delta s_{k\ell}} = \dfrac{\delta}{\delta s_{k\ell}}\left[\dfrac{\delta f_{ij}}{\delta s_{gh}}\right]$

$$= \frac{(2-\delta_{gh})(2-\delta_{k\ell})}{4}\begin{bmatrix} s^{ig}\,s^{hk}\,s^{\ell j} + s^{ig}\,s^{h\ell}\,s^{kj} \\ + s^{hj}\,s^{ik}\,s^{\ell g} + s^{hj}\,s^{i\ell}\,s^{kg} \\ + s^{ih}\,s^{gk}\,s^{\ell j} + s^{ih}\,s^{g\ell}\,s^{kj} \\ + s^{gj}\,s^{ik}\,s^{\ell h} + s^{gj}\,s^{i\ell}\,s^{kh} \end{bmatrix} \qquad (2.5.3)$$

all of which is to be evaluated at

$S = \Gamma = \sum_{\alpha=1}^{2}(n_\alpha - 1)\Omega_\alpha$, or equivalently at:

We note that $\text{cov}(s_{gh}, s_{k\ell})$ cannot be written neatly in terms of the elements of Γ. Without loss of generality, we can make use of the invariance properties of $\theta(\underline{x})$ and it is sufficient to consider only the special case: $\underline{x} \to \underline{x}^* = A\underline{x} + \underline{b}$, with suitable A and $\underline{b}$. In particular,

$$\underline{\mu}_2 \to \underline{\mu}_2 = 0 \qquad \underline{\mu}_1 \to \underline{\mu}_1 \text{ (arbitrary)}$$

$$\Omega_2 \to \Omega_2^* = I_p \qquad \Omega_1 \to \Omega_1^* = \text{diag}(d_1, \ldots d_p)$$

$$\text{(with } d_1 \geq \ldots \geq d_p) \qquad (2.5.4)$$

We now work in this transformed parameter space and drop the superscripts *. For (2.5.2) we have,

$$\sigma_{ij}^{(2)} = \delta_{ij} \text{ and } \sigma_{ij}^{(1)} = d_i\, \delta_{ij}$$

Thus, $\text{cov}(s_{gh}, s_{k\ell}) = (n_1-1)(d_g d_h)(\delta_{gk}\delta_{h\ell} + \delta_{g\ell}\delta_{hk})$

$$+ (n_2-1).1.\ (\delta_{gk}\delta_{h\ell} + \delta_{g\ell}\delta_{hk})$$

where again $\delta_{ij} = \begin{cases} 1 \text{ if } i=j \\ 0 \text{ otherwise.} \end{cases}$

Also, $\Gamma = (n_1-1)\text{diag}(d_1, \ldots, d_p) + (n_2-1)I$

$$= \text{diag}(\gamma_{11}, \ldots \gamma_{pp}) \text{ say, with } \gamma_{ii} = (n_1-1)d_i + (n_2-1)$$

for all i

Thus $\Gamma^{-1} \equiv \text{diag}(\gamma^{11}, .., \gamma^{PP})$ with $\gamma^{ij} = \dfrac{\delta_{ij}}{\gamma_{ii}}$ for all i

Putting together (2.5.2) and (2.5.3) into (2.5.1) in the transformed space, and summing over all g, h, k and ℓ we have,

$$E(s^{ij}) \doteq \gamma^{ij}$$

$$+ \frac{1}{8} \sum_g \sum_h \sum_k \sum_\ell \left[\begin{matrix}(n_1-1)(d_g d_h) \\ + (n_2-1).1\end{matrix}\right] (\delta_{gk}\delta_{h\ell} + \delta_{g\ell}\delta_{hk}) \quad \{\ldots\}$$

where

$$\{\ldots\} = \left[\begin{matrix} \dfrac{\delta_{ig}}{\gamma_{ii}} \left[\dfrac{\delta_{hk}\delta_{\ell j}}{\gamma_{hh}\gamma_{jj}} + \dfrac{\delta_{h\ell}\delta_{kj}}{\gamma_{hh}\gamma_{jj}}\right] + \dfrac{\delta_{ih}}{\gamma_{ii}} \left[\dfrac{\delta_{gk}\delta_{\ell j}}{\gamma_{gg}\gamma_{jj}} + \dfrac{\delta_{g\ell}\delta_{kj}}{\gamma_{gg}\gamma_{jj}}\right] \\ + \dfrac{\delta_{hj}}{\gamma_{jj}} \left[\dfrac{\delta_{ik}\delta_{\ell g}}{\gamma_{ii}\gamma_{gg}} + \dfrac{\delta_{i\ell}\delta_{kg}}{\gamma_{ii}\gamma_{gg}}\right] + \dfrac{\delta_{gj}}{\gamma_{jj}} \left[\dfrac{\delta_{ik}\delta_{\ell h}}{\gamma_{ii}\gamma_{hh}} + \dfrac{\delta_{i\ell}\delta_{kh}}{\gamma_{ii}\gamma_{hh}}\right] \end{matrix}\right]$$

$$\text{i.e. } E(s^{ij}) \simeq \frac{\delta_{ij}}{\gamma_{ii}} + \frac{1}{8}\frac{[(n_1-1)d_i^2 + (n_2-1)1]\delta_{ij}.8}{\gamma_{ii}^3}$$

$$+ \frac{1}{8}\frac{\delta_{ij}}{\gamma_{ii}^2}\sum_{h=1}^{p}\left[\frac{(n_1-1)d_i d_h + (n_2-1)}{\gamma_{hh}}\right] 8$$

since for example:

$\{\delta_{gk}\,\delta_{h\ell}\,\delta_{ig}\,\delta_{hk}\,\delta_{\ell j}\} = 0$ unless $i=g=k=h=\ell=j$, when it is equal to 1.

while,

$\{\delta_{gk}\,\delta_{h\ell}\,\delta_{ig}\,\delta_{h\ell}\,\delta_{kj}\} = 0$ unless $i=g=k=j$ and $h=\ell$ when it is equal to 1

and also since in the expansion there are 8 terms of each type.

We therefore have

$$\underline{E(S^{-1}) \simeq \Gamma^{-1} + \epsilon} \qquad (2.5.5)$$

where $\Gamma = \text{diag}\,(\gamma_{11},..,\gamma_{pp})$, and $\gamma_{ii} = (n_1-1)d_i+(n_2-1)$

and $\epsilon = \text{diag}(\epsilon_{11},..,\epsilon_{pp})$ where

$$\epsilon_{ii} = \frac{1}{\gamma_{ii}^2}\left[\frac{[(n_1-1)d_i^2 + (n_2-1).1^2]}{[(n_1-1)d_i + (n_2-1).1]} + \sum_{h=1}^{p}\frac{[(n_1-1)d_i d_h + (n_2-1).1^2]}{[(n_1-1)d_h + (n_2-1).1]}\right]$$

Note that γ_{ii} is of order $\frac{1}{n}$ and ϵ_{ii} is of order $1/(n^2)$.

Returning to the question of bias, we have: for i=1,2 (and in the transformed parameter space):

$$\alpha_i^2(\underline{x}) = \begin{cases} (\underline{x}-\underline{\mu}_1)^T \text{diag}\left[\frac{1}{d_1},\ldots,\frac{1}{d_p}\right](\underline{x}-\underline{\mu}_1) & \text{for population 1} \\ \underline{x}^T\underline{x} & \text{for population 2} \end{cases}$$

whereas:

$$E(\hat{h}_i^2(\underline{x})) \simeq (n_1+n_2-p-3)(\underline{x}-\underline{\mu}_i)^T(\Gamma^{-1}+\epsilon)(\underline{x}-\underline{\mu}_i)$$

$$+ \left[(n_1+n_2-p-3)\text{trace}\{(\Gamma^{-1}+\epsilon)\text{cov}(\bar{\underline{x}}_i)\} - \frac{p}{n_i}\right]$$

Let $r_i = E(\hat{h}_i^2(\underline{x})) - \alpha_i^2(\underline{x})$

$$= (\underline{x} - \underline{\mu}_i)^T\,[(n_1+n_2-p-3)(\Gamma^{-1}+\epsilon) - \Omega_i^{-1}](\underline{x}-\underline{\mu}_i)$$

$$+ \frac{1}{n_i} [\text{trace } \{(n_1+n_2-p-3)(\Gamma^{-1}+\epsilon)\ \Omega_i\} - p]$$

where Ω_i and $\underline{\mu}_i$ are defined in (2.5.4).

Finally,

$$\text{BIAS1} \doteq \frac{1}{2}\log(|\Omega_2|/|\Omega_1|) + \frac{1}{2}(r_1 - r_2) \qquad (2.5.6)$$

We will check the performance of BIAS1 empirically in Chapter 3.

CHAPTER 3

SIMULATION STUDY

3.1 General description

(A): The need for simulation

Knowledge of the sampling distribution of $\hat{e}$ (the estimated log-odds) would be a useful prerequisite to constructing interval estimates for e. This distribution is not currently known. However, we can still carry out simulations to study this distribution and the performance of approximate interval estimation techniques. The technique of simulation constrains us to look at specific cases of the distributions for the two underlying populations. By a careful selection of the particular values of the parameters we can hopefully cover a wide range of interesting situations.

The different parameters to be considered involve,

(i) sample sizes n_1 and n_2

(ii) dimensionality, p

(iii) equal and unequal covariance matrices

(iv) positions of the population means

(v) the position of the $\underline{x}$ vector

(B): Simulating the empirical distribution of $\hat{e}$

We have, Π_i $\underline{x} \sim Np(\underline{\mu}_i, \Omega)$ (i=1,2).

By using $\underline{x}^{\#} = A\underline{x} + \underline{b}$ for suitable A and $\underline{b}$, plus the fact that e and $\hat{e}$ are invariant under such linear transformations, Critchley and Ford (1985) converted the two population discrimination problem, with $\Omega_1 = \Omega_2$, into the following canonical form:

$$\Pi_i;\ \bar{\underline{x}}_i^{\#} \sim Np(\underline{\mu}_i^{\#}, \frac{1}{n_i}I) \text{ and } S^{\#} \sim Wp(N, I) \qquad (3.1.1)$$

where $\underline{\mu}_1^{\#} = \frac{1}{2}\underline{\delta}$, $\underline{\mu}_2^{\#} = -\frac{1}{2}\underline{\delta}$, $\underline{\delta} = (\Delta, 0, .., 0)^T$

$\Delta^2 = (\underline{\mu}_1 - \underline{\mu}_2)^T \Omega^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$, I = identity matrix

$N = n_1 + n_2 - 2$

For the unequal covariance case with,

Π_i: $\underline{x} \sim Np(\underline{\mu}_i, \Omega_i)$ i=1,2 (see Appendix 3.1),

the canonical form is:

Π_1: $\underline{x}^{\#} \sim Np(\underline{\mu}, D)$ and Π_2: $\underline{x}^{\#} \sim Np(0, I$

where $D = diag(d_1, d_2, \ldots d_p)$, $d_i \geq 0$ i=1,..,p (3.1.2)

$\underline{\mu}$ = arbitrary constant p-vector.

(3.1.2) will be used in this study. Figure [3(i)] is a flow-chart showing the essential steps involved in estimating, for example, the first four moments of the empirical distribution of $\hat{\theta}$, and a brief description is as follows:

(i) generate $\bar{\underline{x}}_1$, $\bar{\underline{x}}_2$, S_1, S_2 (see Appendix 3.2); these are the statistics required to calculate $\hat{\theta}$.

(ii) calculate $\hat{\theta}$, see (2.1.2) and (2.1.4).

(iii) since we know the true mean and covariance matrix, we know the true log-odds, θ_T, for given "patient" vector $\underline{x}$. Let NREPL be the number of $\hat{\theta}$ to be generated. In this study, NREPL = 10,000. Further, let

$$\bar{\theta} = \left[\sum_{i=1}^{NREPL} \hat{\theta}_i\right]/NREPL \quad \{\hat{\theta}_i = i^{th} \text{ replicate of } \hat{\theta} \quad i = 1,..,10000\}$$

Define $\hat{M}(K) = \sum_{i=1}^{NREPL} (\hat{\theta}_i - \bar{\theta})^k$; k=2,3,4 (3.1.3).

Direct computation of M(K) using sums of powers of θ_i and θ may yield inaccurate numerical values. In this study we calculate $\hat{M}(K)$ from the expansion of

$$\sum_{i=1}^{NREPL} [(\hat{\theta}_i - \theta_T) + (\theta_T - \bar{\theta})]^k$$

(C) Choice of parameters

From figure [3(i)], the choice of n_1, n_2, p, $\underline{\mu}$, D, $\underline{x}$ will determine the distribution of $\hat{\theta}$. We shall use the term

READ-IN

(i) patient vector, $\underline{x}$

(ii) TRUE LOG-ODDS = θ_T

(iii) Known parameters, viz:-
n_1, n_2, p
$\underline{\mu}$, D
NREPL (=10,000)

I=1
SUM1=0, SUM2=0
SUM3=0, SUM4=0

Generate Data, viz:-

$\underline{\bar{x}}_1^{\#} \sim Np(\underline{\mu}, \frac{1}{n_1} D)$, $S_1^{\#} \sim Wp\{n_1-1, D\}$

$\underline{\bar{x}}_2^{\#} \sim Np(\underline{0}, \frac{1}{n_2} I)$, $S_2^{\#} \sim Wp\{n_2-1, I\}$

Calculate

$\hat{\theta}_I (\underline{\bar{x}}_1^{\#}, \underline{\bar{x}}_2^{\#}, S_1^{\#}, S_2^{\#} | \underline{x})$

I denote replication

SUM1 = SUM1 + $\hat{\theta}_I$

SUM2 = SUM2 + $(\hat{\theta}_I - \theta_T)^2$

SUM3 = SUM3 + $(\hat{\theta}_I - \theta_T)^3$

SUM4 = SUM4 + $(\hat{\theta}_I - \theta_T)^4$

IS I=NREPL

NO

I=I+1

YES

$\bar{\theta}$ = SUM1/NREPL

BIAS = $\theta_T - \bar{\theta}$

$\hat{M}(2)$ = SUM2 - NREPL*BIAS^2

$\hat{M}(3)$ = SUM3 + 3BIAS*SUM2-2NREPL*BIAS^3

$\hat{M}(4)$ = SUM4 + 4BIAS*SUM3+6BIAS^2*SUM2
- 3NREPL*BIAS^4

STOP

FIGURE (3(i)): Calculating moments of $\hat{\theta}$

"SIMULATION", henceforth, to mean the particular values (fixed) of n_1, n_2, p, $\underline{\mu}$, D, $\underline{x}$ that were used to generate the ten thousand (= NREPL) $\hat{e}_l$(l=1,...,10000). Hence, we have 10,000 independent REPLICATES of $\hat{e}$ for each simulation.

We list the values of the parameters considered. Each simulation will be some combination of these values.

(i) n_i = 400, 40, 20 (i=1,2)

n_i = 400; 'would' represent asymptotic sample sizes.

n_i = 40; is a moderate sample size.

n_i = 20; represents small sample situations. We considered special cases of $n_1=n_2$, $n_1>n_2$, $n_1<n_2$.

(ii) p = 2; for the simple case

p = 5; to get a feel of complex situations in higher dimensions.

(iii) D = diag $\{d_1,\ldots,d_p\}$

Firstly considered $\Omega_1=\Omega_2$, i.e. $d_i=1$ $\forall_i$.

Also looked at $d_i=d$ where d=4 and 0.25.

Special cases of $d_{i+1}<d_i$ were studied (for both p=2 and p=5).

(iv) $\underline{\mu}^T = (\mu_1, \mu_2, \ldots, \mu_p)$

The study is mostly for $\mu_1=2$ and $\mu_j=0$ (j=2,...,p).

For p=2, looked at $\underline{\mu}^T = (\sqrt{2}, \sqrt{2})$ and $\underline{\mu}^T = (0,2)$.

For p=5, considered a special case where $\underline{\mu}^T = (y,y,y,y,y)$ with $y=2/\sqrt{5}$.

(v) Nearly all simulations used $\Delta = 2$. This means greater overlap of the two populations when p=5. As a special case we looked at $\Delta = 6.07$ for p=2. This special case is equivalent to having a probability of misclassification of 0.001 if we assume $\Omega_1=\Omega_2$ and equal 'prior' probabilities.

(vi) "patient"-vector $\underline{x}$;

We use the result that for $f(\underline{x}|\Pi_i) \sim N_p(\underline{\mu}_i, \Omega_i)$,

then $(\underline{x}-\underline{\mu}_i)^T \Omega_i^{-1}(\underline{x}-\underline{\mu}_i) \sim \chi^2(p)$. (i=1,2) (3.1.4)

There is, by (3.1.4), a probability of, say, 0.90 of being within an ellipsoid, which for convenience we shall refer to as a "probability ellipsoid".

Define \$(C2%, C1%) as an $\underline{x}$-point that lies on the intersection of the C2% and C1% probability ellipsoids of population two and population one respectively. We now select the following $\underline{x}$-points:

(a) $\underline{x}^T$ = \$(90%, 90%) ≡ \$9090

(b) $\underline{x}^T$ = \$(90%, 38%) ≡ \$9038

(c) $\underline{x}^T$ = origin

(d) $\underline{x}^T = \underline{\mu}^T$

(e) $\underline{x}^T$ = a point on the x_1-axis, such that

$\underline{x}^T$ = \$(C%, C%) = \$CC

(c) and (d) are 'typical' observations for the relevant population. (a) is regarded as atypical to both populations. (b) is more atypical to population two. (e) in practice will be a 'patient' that would be difficult to assign to either population.

If \$9090 does not exist, we shall replace it by the next nearest point, say \$9085.

Some points, say \$9038, may not be unique in the SIMULATIONS. This will not matter if the d_i's [$D=\text{diag}(d_1, \ldots, d_p)$] are equal since the $\underline{x}$-points corresponding to \$9038 are symmetric about the line of centres. Due to the invariance properties of $\hat{e}(\underline{x})$, the distribution of $\hat{e}(\underline{x})$ is the same for all such $\underline{x}$-points. When the d_i's are unequal for p=2 as long as $\underline{\mu}^T = (\mu, 0)$ it does not matter if the point such as \$9038 is not unique. However, for SIMULATIONS SC(1,1), SC(1,2) and SG(1,1) the points chosen for \$9038 and/or \$9090 clearly are not unique and for convenience were chosen to be

SA; [p=2; $\underline{\mu}^T$=(2,0); Δ=2.0]

	n_1=400=n_2	n_1=40=n_2	n_1=20=n_2	n_1=20 n_2=40	n_1=40 n_2=20
d_1=4=d_2	SA(1,1)	SA(1,2)	SA(1,3)	SA(1,4)	SA(1,5)
d_1=1=d_2	SA(2,1)	SA(2,2)	SA(2,3)	SA(2,4)	SA(2,5)
$d_1=\frac{1}{4}=d_2$	SA(3,1)	SA(3,2)	SA(3,3)	SA(3,4)	SA(3,5)

Simulation(s) SA: Effect on sample sizes for equal and unequal covariance matrices.

SB; [p=2; $\underline{\mu}^T$=(2,0); Δ=2.0; n_1=40; n_2=20]

d_2 \ d_1	4.00	1.00	0.25
4.00	SB(1,1)	-	-
1.00	SB(2,1)	SB(2,2)	-
0.25	SB(3,1)	SB(3,2)	SB(3,3)

Simulation(s) SB: Effect of unequal d_i

Note: SB(1,1) ≡ SA(1,5); SB(2,2) ≡ SA(2,5),

SB(3,3) ≡ SA(3,5)

SC; [p=2, n_1=40, n_2=20, d_1=4, $d_2=\frac{1}{4}$, Δ=2.0]

$\underline{\mu}^T$	($\sqrt{2}$, $\sqrt{2}$)	(0,2)
simulation	SC(1,1)	SC(1,2)

Simulation(s) SC: Effect of $\underline{\mu}^T$

Table (3(i)): Labels to identify simulations

SD(1,1) ≡ [p=2; $n_1=20=n_2$; $d_1=1=d_2$; $\underline{\mu}^T=(\Delta,0)$; $\Delta=6.07$]

Simulation SD: Effect of Δ

SE; [p=5; $\underline{\mu}^T = (2,0,0,0,0)$; $\Delta=2$]

	$n_1=40=n_2$	$n_1=20=n_2$
$d_i=4$ $i=1,\ldots,5$	SE(1,1)	SE(1,2)
$d_i=1$ $i=1,\ldots,5$	SE(2,1)	SE(2,2)
$d_i=0.25$ $i=1,\ldots,5$	SE(3,1)	SE(3,2)

Simulation(s) SE: Effect of sample sizes and particular Ω_1, Ω_2 in higher dimension.

SF(1,1) ≡ [p=5, $\underline{\mu}^T = (2,0,0,0,0)$, $\Delta=2$, $n_1=40=n_2$; D=diag{kz, kz^2, kz^3, kz^4, kz^5}; where k=12, z=0.25]

SF(1,2) ≡ [same as SF(1,1) except; D=diag{2, $^4/_3$, 1.0, 0.75, 0.50}]

Simulation(s) SF: Some special values of d_i, i=1,...,5

SG(1,1) ≡ [p=5, $\Delta=2$, $n_1=40=n_2$; D=diag{2, $^4/_3$, 1, $^3/_4$, $^1/_2$}; $\underline{\mu}^T=(\gamma,\gamma,\gamma,\gamma,\gamma)$; $\gamma=2/\sqrt{5}$]

Simulation SG: Effect of $\underline{\mu}^T$

Table (3(i)) (contd.)

in the positive quadrant of the $\underline{x}$-space with $\underline{x}$ of the form $(x_1, x_2, 0, 0, 0)^T$ in the p=5 case.

For convenience and consistency, we introduce "labels" to denote particular simulations, as given in Table [3(I)]. As an example, the (1,1) element of simulation SA is,

$$SA(1,1) \equiv \left[\begin{array}{l} n_1 = 400 = n_2,\ p=2 \\ D = \text{diag}(4,4),\ \underline{\mu}^T = (2,0),\ \Delta=2.0 \\ \text{"for given } \underline{x}\text{-vector"}. \end{array}\right]$$

(D) Assumptions on Ω_i (i=1,2)

We shall look at the empirical distribution of $\hat{\theta}$ given:

(i) always assume $\Omega_1 = \Omega_2$, i.e. $\hat{\theta} = \hat{\theta}_E$

(ii) always assume $\Omega_1 \neq \Omega_2$, i.e. $\hat{\theta} = \hat{\theta}_{NE}$

(iii) Based on the outcome of the test (see Appendix (4.3))

$$H_O: \ \Omega_1 = \Omega_2 \ \ V_s \ H_1: \ \Omega_1 \neq \Omega_2 \qquad (3.1.5)$$

let $\hat{\theta} = \hat{\theta}_E$ if accept H_O (null hypothesis)

or $\hat{\theta} = \hat{\theta}_{NE}$ if reject H_O.

We can think of procedure (i) as STAT1, a statistician who always uses the Linear discriminant function. Likewise (ii) is a statistician, STAT2, who will only use the quadratic discriminant function. Procedure (iii) is STAT3 a statistician whose decision will depend on the outcome of the test of equality of covariance matrices.

For convenience we shall henceforth refer to the test in (3.1.5) simply as TEST. To investigate the performance of the TEST, define

P_O = probability reject the null hypothesis (i.e. H_O) of TEST.

In each simulation we generate 10,000 replicates of $\hat{\Omega}_1$ and $\hat{\Omega}_2$.

We therefore have an estimate of P_O, where,

$\hat{P}_O$ = proportion of 10,000 results of TEST when we reject the null hypothesis.

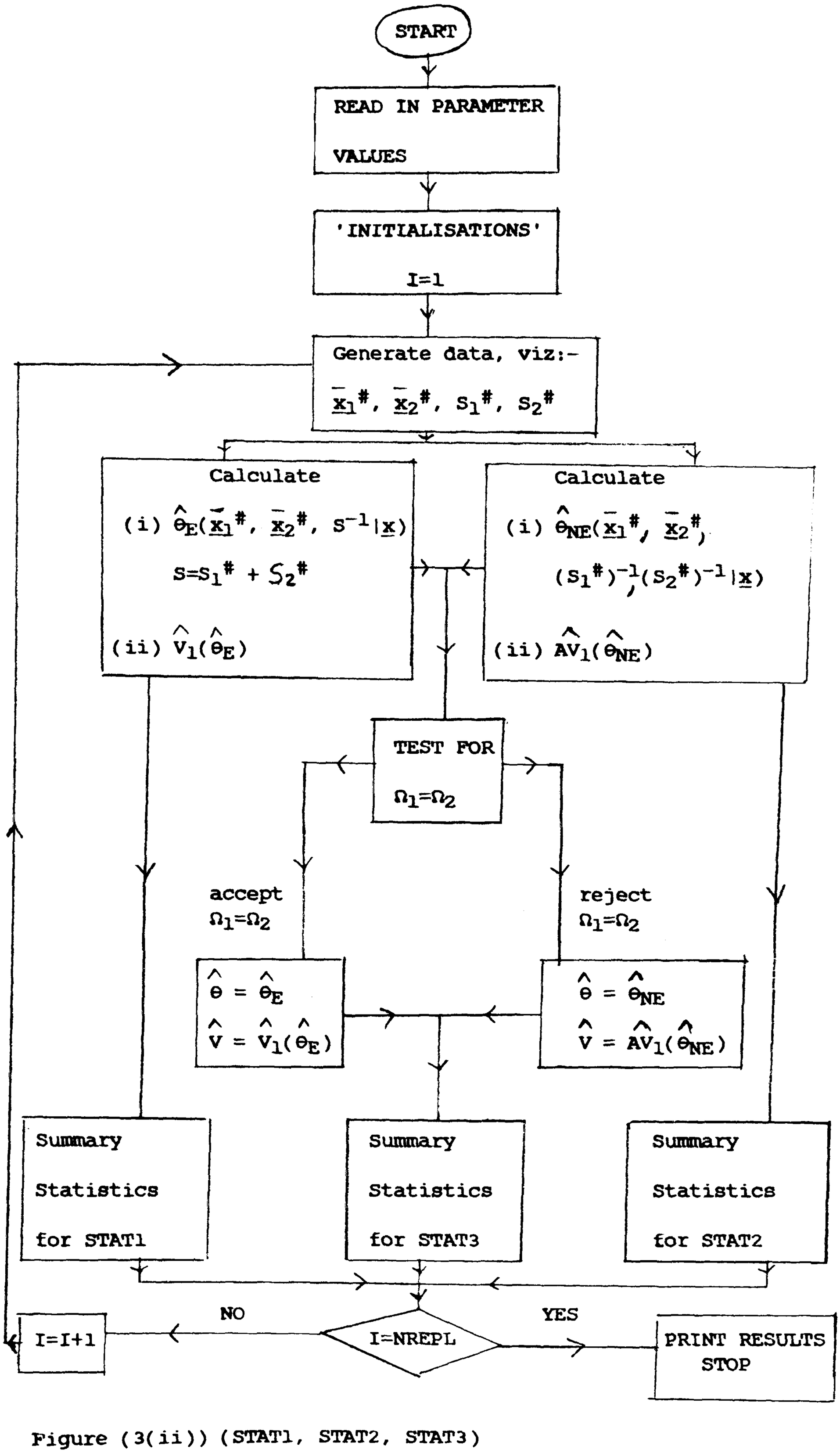


Figure (3(ii)) (STAT1, STAT2, STAT3)

Figure (3(II)) gives an illustration, and summary statistics here include the moments of $\hat{e}$.

(3.2) Performance criteria and other Summary Statistics

(A) Confidence probability

For each SIMULATION (see section 3.1.C) we generate 10,000 replicates of $\hat{e}$. In each replication, we constructed the intervals.

$$(\hat{\theta}_E \pm 1.96 \sqrt{[\hat{V}_1(\hat{\theta}_E)]}) \qquad (3.2.1)$$

$$(\hat{\theta}_{NE} \pm 1.96 \sqrt{[\hat{AV}_1(\hat{\theta}_{NE})]}) \qquad (3.2.2)$$

Define the 'confidence' probabilities: where

CP1 = probability of capturing the true log-odds, θ_T, using (3.2.1)

CP2 = probability of capturing θ_T using (3.2.2).

Further, given the result of TEST (3.1.5),

$$\text{Let } y_i = \begin{cases} 1 & \text{if accept } H_0 \text{ and (3.2.1) capture } \theta_T \\ 1 & \text{if accept } H_1 \text{ and (3.2.2) capture } \theta_T \\ 0 & \text{otherwise} \end{cases}$$

Let CP3 = probability STAT3 captures θ_T, where,

$$\hat{CP3} = \sum_{i=1}^{10000} y_i/10,000$$

We will refer to CP1, CP2, CP3 as the confidence probabilities of STAT1, STAT2, STAT3 respectively (see also Section (3.1.D)).

(B) Zero probabilities

Let Z1 be the probability (3.2.1) captures zero

Let Z2 " (3.2.2) " "

Given the result of TEST (3.1.5),

$$\text{Let } t_i = \begin{cases} 1 \text{ if accept } H_0 \text{ and (3.2.1) capture zero} \\ 1 \quad " \quad H_1 \text{ and (3.2.2)} \quad " \quad " \\ 0 \text{ otherwise} \end{cases}$$

Let Z3 = probability STAT3 captures zero, where,

$$\hat{Z3} = \sum_{i=1}^{10000} t_i/10,000$$

Z1, Z2, Z3 will be referred to as the probability of capturing zero for STAT1, STAT2, STAT3 respectively.

(C) Accuracy of $\hat{CP}$ and $\hat{Z}$

Without loss of generality, consider $\hat{CP}2$. Let Y be the total number of events "(3.2.2) captures θ_T" with probability CP. For each of our SIMULATION,

$$Y \sim \text{Binomial}(10,000,\ CP)$$

$$\text{Thus Var}(\hat{CP}) = CP(1-CP)/10,000$$

and the largest value of this variance is when CP = 0.5. The largest value for the standard deviation of $\hat{CP}$ is 0.005. Therefore $\hat{CP}$, at the very worst, will be accurate to +/- 1 on the second decimal place.

In particular, STAT2 is said to perform better than STAT1 in terms of the confidence probability if $\hat{CP}2$ is "closer" to 0.95 than $\hat{CP}1$.

The 'worst case' accuracy of $\hat{Z}$ is clearly the same as for $\hat{CP}$.

(D) Moments of the distribution of $\hat{\theta}$

In a given simulation we calculate $\bar{\theta}$ (see (3.1.3)). Since we know the true log-odds θ_T, define the estimated Bias as,

$$\hat{\epsilon} = \theta_T - \bar{\theta} \qquad (3.2.3)$$

Further define $\epsilon 1$, $\epsilon 2$, $\epsilon 3$ as the bias for STAT1, STAT2, STAT3 respectively, where $\bar{\theta}$ is appropriately calculated for the relevant statistician in (3.2.3).

Using the notation of section (3.1.B) define,

$$\hat{V}_S(\hat{\theta}_{NE}) = \frac{[\hat{M}(2) \text{ for STAT2}]}{(NREPL-1)} \qquad (3.2.4)$$

where $\hat{M}(2)$ is equation (3.1.3) for STAT2. The subscript 's' in $\hat{V}_S(\hat{\theta}_{NE})$ serves to indicate that it is the estimated sampling variance of $\hat{\theta}_{NE}$ for a given SIMULATION. By calculating the corresponding $\hat{M}(2)$ for STAT1, we have

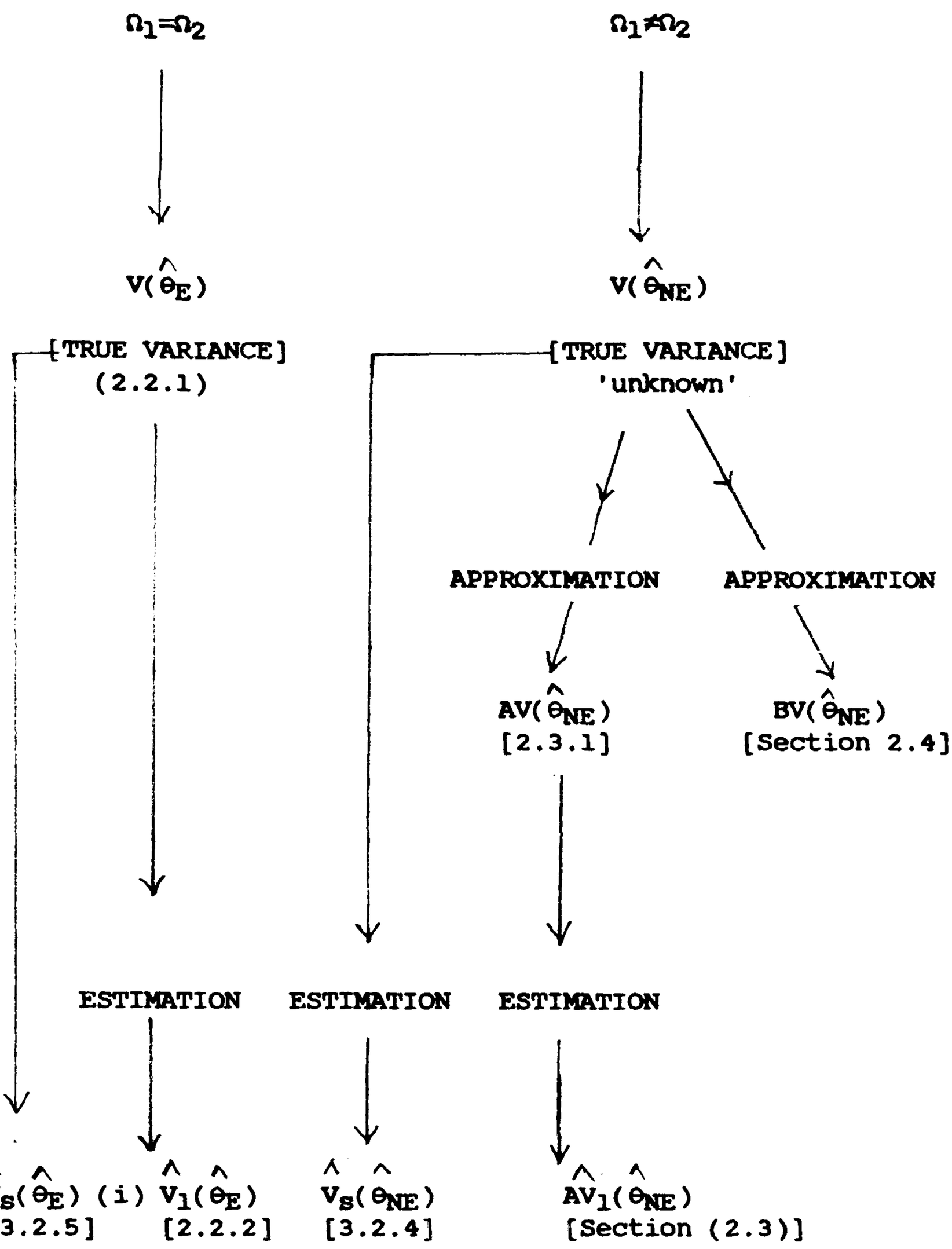


Figure [3(iii)] Summary of variance formulae.

Note: (1) $\hat{\theta}_E$ and $\hat{\theta}_{NE}$ are $\hat{\theta}_E(\underline{x})$ and $\hat{\theta}_{NE}(\underline{x})$ respectively.

(2) Below each variance is the relevant reference.

$$\hat{V}_s\,(\hat{\theta}_E) = \frac{[\hat{M}(2) \text{ for STAT1}]}{(\text{NREPL}-1)} \qquad (3.2.5)$$

See Figure [3(III)] for a summary of the formulae used for the variance of $\hat{\theta}$.

Although we have the exact Variance of $\hat{\theta}_E(\underline{x})$, we will use (3.2.5) when we have $\Omega_1 \neq \Omega_2$

In section (3.1.D), depending on the result of the TEST, define

$$\hat{\theta}_3(\underline{x}) = \begin{cases} \hat{\theta}_E(\underline{x}) \text{ if accept } H_0 \text{ of TEST} \\ \hat{\theta}_{NE}(\underline{x}) \text{ if accept } H_1 \text{ of TEST} \end{cases}$$

Define,

$$\hat{V}_s\,(\hat{\theta}_3) = \frac{[\hat{M}(2) \text{ for STAT3}]}{\text{NREPL}-1} \qquad (3.2.6)$$

where $\hat{M}(2)$ is equation (3.1.3) based on the empirical distribution of $\hat{\theta}_3(\underline{x})$.

As measures of non-normality we define, $\beta \equiv$ skewness and $\gamma \equiv$ kurtosis whose estimates are,

$$\hat{\beta} = \hat{T}(3)/[\hat{T}(2)]^{3/2}$$

$$\hat{\gamma} = \hat{T}(4)/[\hat{T}(2)]^{2}$$

where $\hat{T}(K) = \hat{M}(K)/(\text{NREPL}-1)$ and where $\hat{M}(K)$ is equation (3.1.3) for the appropriate statistician (STAT1, STAT2 or STAT3).

If $\hat{\theta}$ has a normal distribution,

$$\hat{\beta} \dot{\sim} N(0,\ {}^{6}/_{\text{NREPL}})$$

and,

$$\hat{\gamma} \dot{\sim} N(3,\ {}^{24}/_{\text{NREPL}})$$

See for example Snedecor and Cochran (1967) page 86 and 87. Since NREPL equals 10,000 we will regard the values of $\hat{\beta}$ outside the interval (0 ± 0.05) as showing significant asymmetry in the distribution of $\hat{\theta}$. Similarly, we have significant kurtosis if $\hat{\gamma}$ is outside the interval (3 ± 0.10).

Further define,

$\hat{\beta}1$, $\hat{\beta}2$, $\hat{\beta}3$ as $\hat{\beta}$ for STAT1, STAT2, STAT3 respectively.

$\hat{\gamma}1$, $\hat{\gamma}2$, $\hat{\gamma}3$ as $\hat{\gamma}$ for . . . .

(3.3) Performance of $AV(\hat{\theta}_{NE})$ and $V(\hat{\theta}_E)$

(A) Performance of $AV(\hat{\theta}_{NE})$

Since NREPL is large, we would expect $V_S(\theta_{NE})$ to be close to the true variance of $\hat{\theta}_{NE}$, $V(\hat{\theta}_{NE})$ [see Appendix (3.3)].

We will compare $AV(\hat{\theta}_{NE})$ [equation (2.3.1)] with $\hat{V}_S(\hat{\theta}_{NE})$, and investigate the effect of sample sizes and dimensionality (simulations SA and SE). TABLE (3(II)) shows some results.

For p=2, both variances are almost identical, but differ more for small, and unequal, sample sizes. Generally both variances increase as the n_i decrease. A combination of large p and small n_i can also increase both variances, e.g. for $\underline{x}^T$ = \$CC in SA(1,3) and SE(1,2).

The difference between $AV(\hat{\theta}_{NE})$ and $\hat{V}_S(\hat{\theta}_{NE})$ is small. The direction of this difference depends critically on the $\underline{x}$-point and/or value of (d_1, d_2) chosen.

The results so far suggest that for large sample sizes $AV(\hat{\theta}_{NE})$ may be well estimated. We may however have a problem in estimating $AV(\hat{\theta}_{NE})$ for small sample situations and/or when p>2.

(B) Comparing $V(\hat{\theta}_E)$ and $V(\hat{\theta}_{NE})$ when $\Omega_1=\Omega_2$

Since θ_{NE} involves estimation of more parameters (unnecessarily in this case), we would expect that $V(\hat{\theta}_{NE})$ will typically be larger than $V(\hat{\theta}_E)$. Again using $\hat{V}_S(\hat{\theta}_{NE})$ to estimate $V(\hat{\theta}_{NE})$, we shall compare $\hat{V}_S(\hat{\theta}_{NE})$ with $V(\hat{\theta}_E)$ [see figure 3(III) for summary of Notation used for the variance of $\hat{\theta}$]. The results are given in Tables (3(III)a), (3(III)b), ... , (3(III)g). We look at the last column for the simulations SA(2, j); j=1,5 and SE(2, ℓ); ℓ=1,2. In all cases both variances increase when sample sizes decrease and

$\hat{V}_S(\hat{\theta}_{NE})$ is constantly greater than $V(\hat{\theta}_E)$. However, even for $n_1=20=n_2$, $p=2$, $\hat{V}_S(\hat{\theta}_{NE})$ is never greater than 1.5 times that of $V(\hat{\theta}_E)$.

The effect of increasing p to 5 is again to increase both variances, but the increase is larger for $\hat{V}_S(\hat{\theta}_{NE})$.

(3.4) Comparing STAT1 and STAT2

In the SIMULATIONS SA(i,j); all (i,j) and SE(k,ℓ); all (k,ℓ), the estimated probability of rejecting H_o of TEST, i.e. $\hat{P}_o$, (see Section (3.1.D)) is either "close" to 0.95 or 0.05 and by definition STAT3 is essentially "similar" to STAT2 or STAT1. Therefore in the simulations SA(i,j); all (i,j) and SE(k,ℓ); all (k,ℓ), we shall in this section only report;

θ_T, $\hat{CP}1$, $\hat{\epsilon}1$, $\hat{V}_S(\hat{\theta}_E)$ or $V(\hat{\theta}_E)$

and $\hat{P}_o$, $\hat{CP}2$, $\hat{\epsilon}2$, $\hat{V}_S(\hat{\theta}_{NE})$.

When $\Omega_1=\Omega_2$, $\hat{P}_o$ is close to the nominal significance level of 0.05 but when $\Omega_1\neq\Omega_2$, $\hat{P}_o$ frequently exceeds the value of 0.95. For example in SA(1,1) we have $\hat{P}_o$ equal to one. We feel that these values of $\hat{P}_o$ merit the exclusion of STAT3 in this section.

When $\Omega_1=\Omega_2$, $\hat{CP}1$ is similar to $\hat{CP}2$ for all simulations with values of 0.95 to 1.00. Both Statisticians have unbiased estimates of θ_T and are 'correct' in their analyses, therefore both having very similar $\hat{CP}$ values close to the target value of 0.95. For smaller sample sizes and/or p=5, $\hat{CP}2$ tends to be larger than $\hat{CP}1$, possibly related to the accuracy with which $V(\hat{\theta}_{NE})$ and $V(\hat{\theta}_E)$ are estimated and the validity of the assumed approximate normal distributions for $\hat{\theta}$.

When $\Omega_1\neq\Omega_2$, $\hat{CP}1$ can be anything from zero to 0.99, while $\hat{CP}2$ takes values of 0.93 to 1.00. This is due to

(i) STAT2 is making correct assumptions and STAT1 is not.

(ii) STAT1 has a biased estimate of θ_T

(iii) $V(\hat{\theta}_E)$ tends to be smaller or not very different from $V_s(\hat{\theta}_{NE})$. This suggests that the interval in (3.2.1) is generally shorter than that in (3.2.2); and because of (ii) STAT1 clearly could have less chances of "capturing" θ_T.

For instance, looking at say SIMULATION SE(1,2) where $n_1=20=n_2$ and p=5, $\hat{CP1}$ is either equal or close to zero and we note that $\hat{\epsilon 1}$ is large. Clearly the bias in estimating θ_T is an important feature.

The simulations in these sections generally showed CP2 tends to be larger than the target value of 0.95 with $\hat{CP2}$ frequently taking values of 0.97 and 0.98 for small sample situations.

(3.5) Special cases of unequal covariance matrices, the mean vector $\underline{\mu}$ and Δ

In this section the results are given in Tables (3(iv)a), (3(iv)b),....,(3(iv)e).

(A) Effect of $d_{i+1} > d_i$ in $D=diag(d_1, d_2, \ldots, d_p)$

We consider the SIMULATIONS SB(2,1), SB(3,1), SB(3,2), SF(1,1) and SF(1,2). In these simulations $\hat{P}_o$ may well be below 0.95 and not close to 0.05. Only for such values of $\hat{P}_o$ will we also report $\hat{CP3}$, $\hat{\epsilon 3}$ and $\hat{V}_s(\hat{\theta}_3)$, so as to compare STAT3 with either of the other two Statisticians.

STAT2's performance relative to STAT1 is essentially the same here as in Section (3.4) and in particular $\hat{CP2}$ tends to be above the value 0.95. As before STAT1 has poorer $\hat{CP}$ values mainly because of his biased estimates of θ_T. STAT1's bias, that is $\hat{\epsilon 1}$, does depend on the values of $d_i (i=1,\ldots,p)$. For example, SF(1,1) has larger $\hat{\epsilon 1}$ values than SE(1,1) and SE(3,1), while SF(1,2) has smaller $\hat{\epsilon 1}$ values than SE(1,1) and SE(3,1) suggests that the more different the values of a particular set of d_i (i=1,....,5) are, the

more biased STAT1 will be in estimating θ_T.

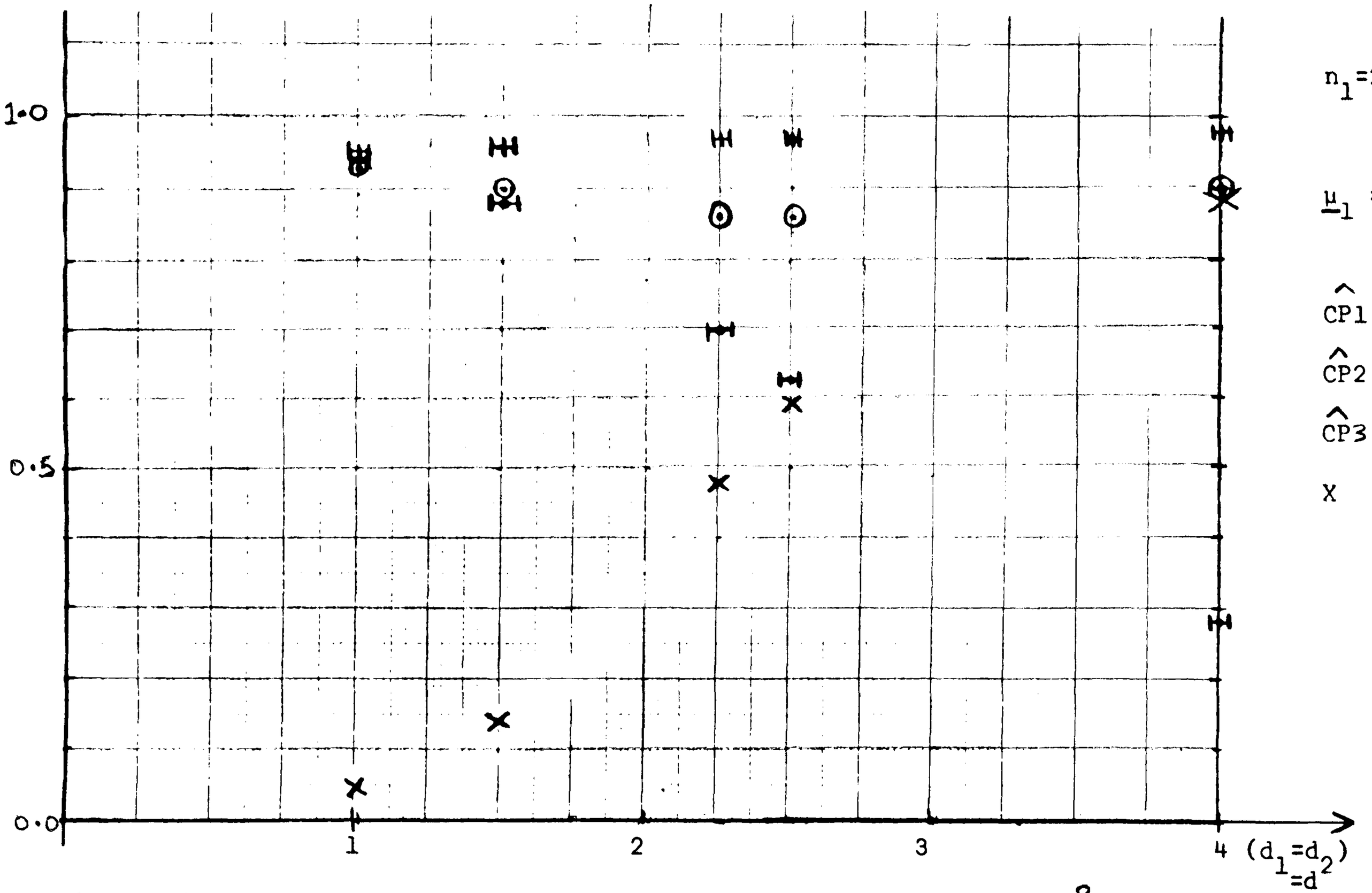
STAT3's performance depends critically on the value of $\hat{P}_O$. In SB(2,1) and SB(3,2) the values of $\hat{P}_O$ are 0.80 and 0.84, while for SF(1,2), $\hat{P}_O$ equals 0.47. Note that $\hat{V}_S(\hat{\theta}_3)$ is often larger than $\hat{V}_S(\hat{\theta}_E)$ and $\hat{V}_S(\hat{\theta}_{NE})$. It is difficult to explain why this is happening but it is clearly due to the "DUAL IDENTITY" of STAT3 resulting in a very complicated distribution of $\hat{\theta}_3(\underline{x})$. Further note that in cases where the power to detect $\Omega_1 \neq \Omega_2$ is not close to one, the non-equality of variances (i.e. unequal d_i's) can still have a substantial effect on STAT3's confidence probability. Hence the recommendation "test $\Omega_1=\Omega_2$ first and act accordingly" may be a dangerous activity since we may not have sufficient power to detect differences in covariance structure which can substantially affect STAT1's performance. Admittedly STAT3 appears in only a few of our SIMULATIONS and all remarks made about STAT3 should be regarded with caution.

(B) Effect of $\underline{\mu}$

We consider the simulations SC(1,1), SC(1,2) and SG(1,1). For SC(1,1) and SC(1,2), $\hat{P}_O$ equals 0.99. Changing $\underline{\mu}$ can considerably increase $\hat{\epsilon}1$, e.g. for the origin in SB(3,1) and either of SC(1,1) and SC(1,2). The relationship between STAT1 and STAT2 is otherwise similar to that in section (3.4).

For SG(1,1), $\hat{P}_O$ equals 0.46 and comparisons between all Statisticians is similar to SF(1,2), in particular STAT3's performance depends on how "badly" STAT1 performs.

We note that $\underline{x}$-points such as \$9090 and \$9038 are not unique in these SIMULATIONS due to the change of $\underline{\mu}$ and unequal variances (i.e. d_i's). This has been pointed out in section (3.1.C) and we note here the choice of such $\underline{x}$-points in the positive quadrant of the $\underline{x}$-space is simply for convenience.



$n_1=20=n_2$, $p=2$

$\underline{\mu}_1 = \begin{pmatrix}2\\0\end{pmatrix}$, $\underline{\mu}_2 = \begin{pmatrix}0\\0\end{pmatrix}$, $\underline{x} = \begin{pmatrix}0\\0\end{pmatrix}$

$\hat{CP}1 \equiv$ ⊢●⊣

$\hat{CP}2 \equiv$ ⊢⊣

$\hat{CP}3 \equiv$ ⊙

X ≡ $\hat{P}_o$

Figure 3(iv): What sort of values of $d_1=d=d_2$ before STAT1 'useless'?

(C) Effect of Δ

Our study so far considers the case where there is fairly substantial "over-lap" between the two populations. In response to a query in a seminar we looked at a case, i.e. SD(1,1), where the two populations are separated such that the 99% probability ellipsoids (see (3.1.4)) of both populations just touch.

We compare SD(1,1) with SA(2,3). We note that here $\Omega_1=\Omega_2$ for both simulations, and having larger e's and associated variances in SD(1,1) does not dramatically change $\hat{CP}1$ and $\hat{CP}2$. Therefore we can say, with some caution, that at least for the equal covariance case we do not expect the amount of "overlapping" between populations to effect our general conclusions in any dramatic way.

(3.6) $\hat{CP}$, $\hat{P}_o$ and values of d_i $(i=1,\ldots,p)$.

The simulation SF(1,2), with P_o approximately equal to 0.47 suggests that the TEST (see (3.1.5)) may not be sensitive enough to pick out values of $d_i(i=1,\ldots,p)$ close to one. We consider a special set of SIMULATIONS, comparing $\hat{CP}$ and $\hat{P}_o$ values over a number of values of $d_1=d=d_2$. The details are given in figure (3(iv)). Clearly the power of the TEST depends on the values of d, for example when d=2.25, $\hat{P}_o$ is only 0.48. Obviously the performance of STAT3 with respect to $\hat{CP}$ values depends on the power of the TEST. This is shown in figure (3(iv)) where we can see that the $\hat{CP}3$ values are "pulled" downwards by the poor performance of STAT1. The $\hat{CP}3$ values for simulations in earlier sections suggest STAT3's performance may be worse when p=5.

(3.7) Skewness and Kurtosis

There is evidence that $\hat{\beta}$ and $\hat{\gamma}$ are significantly different from their normal distribution values and in some cases are large. However, with respect to $\hat{CP}$ values, the bias and variance of the

distribution of $\hat{\theta}$ are more relevant than having significant $\hat{\beta}$ and $\hat{\gamma}$. We illustrate with $\hat{CP}1$ when $\Omega_1=\Omega_2$ (Table 3(v)a) and with $\hat{CP}2$ when $\Omega_1\neq\Omega_2$ (Table (3(v)b). Clearly, $\hat{\beta}$ and $\hat{\gamma}$ have little effect on $\hat{CP}$ when $\hat{\theta}$ is unbiasedly estimated.

When $n_1=n_2$, $\Omega_1=\Omega_2$, θ_T=zero, Table (3(v)a) shows that $\hat{\beta}1$ and $\hat{\beta}2$ are non-significant. Of course in these cases $\beta1$, $\beta2$ will be exactly zero due to symmetry arguments. When $\theta_T\neq$zero the distribution of $\hat{\theta}$ may be asymmetric.

Generally increasing sample sizes for a given $\underline{x}$-point in Tables (3(v)a), (3(v)b) and (3(v)c) decreases β, γ as expected and improves $\hat{CP}$. Within a set of SIMULATIONS, say SA(1,1), relating results for $\hat{CP}$ with β, γ for different $\underline{x}$-points provides no general conclusions. The difficulty in relating $\hat{CP}$ with β and γ is because $\hat{CP}$ also depends on our ability to estimate $Var(\hat{\theta})$ and because $Var(\hat{\theta})$ will be correlated with $\hat{\theta}$.

(3.8) Bias of $\hat{\theta}_E(\underline{x})$ when $\Omega_1\neq\Omega_2$

We have seen in earlier sections that for the special cases of $\Omega_1\neq\Omega_2$, STAT1 performed poorly with respect to his confidence probability. The interval estimate $\hat{\theta}_E(\underline{x}) \pm 1.96\sqrt{Var(\hat{\theta}_E(\underline{x}))}$ frequently does not capture θ_T because:

(i) $\hat{\theta}_E(\underline{x})$ is biased, possibly badly biased when $\Omega_1\neq\Omega_2$

(ii) $Var\{\hat{\theta}_E(\underline{x})\}$ is generally small, i.e. STAT1 has a narrower interval.

It is therefore useful to know the bias of $\hat{\theta}_E(\underline{x})$ in situations where there is evidence suggesting unequal covariance matrices. In Chapter 2, an approximate bias of $\hat{\theta}_E(\underline{x})$ when $\Omega_1\neq\Omega_2$ is given in (2.5.6). For purposes of notation we will denote this bias as BIAS1. We will compare $\hat{\epsilon}1$(see(3.2.3)) with BIAS1, and note that the former should be close to the true bias. The values of $\hat{\epsilon}1$ are taken from the simulations SA(i,j), i=1 and 3, j=1,2,3; and

SE(1,1), SE(1,2). TABLE (3(vi)) gives the values of $\hat{\epsilon}_1$ and BIAS1, the former in parenthesis.

When $n_1=400=n_2$ in SA(1,1) and SA(3,1) both bias are almost identical, showing BIAS1 to be a good approximation of the true Bias.

For SA(1,j): j=1,2,3 i.e. when $\Omega_1=4\Omega_2$, $\Omega_2=I$, even for $n_1=20$ (i=1,2) the approximation for BIAS1 is still good. However for SA(3,j): j=1,2,3 when $\Omega_1=\Omega_2/4$, reducing sample size to 20 gives a steadily poorer approximation.

For SE(1,1) and SE(1,2), increasing dimensionality to 5 (p=5) does not make the approximation very different from the p=2 case. In Tables (3(iii)a),....(3(iii)g), (3(iv)a),....(3(iv)e), for large values of $\hat{\epsilon}_1$, in particular those greater than three, $\sqrt{\hat{V}_S}(\hat{\theta}_{NE})$ tends to be about 1.5 or 2 times the size of $\sqrt{\hat{V}_S}(\hat{\theta}_E)$. Clearly the empirical distribution of $\theta_E(\underline{x})$ is more "tightly" packed around it's mean, suggesting that the bias in these situations is well approximated. It is therefore possible that the approximation to the bias may not depend on p but on the size of the bias itself.

We note that only a few cases are considered here.

(3.9) Performance of $BV(\hat{\theta}_{NE}(\underline{x}))$

In Section (2.4.C) we have another approximation to the variance of $\hat{\theta}_{NE}(\underline{x})$; viz:- $BV(\hat{\theta}_{NE}(\underline{x}))$. Table 3(vii) gives some comparison of $AV(\hat{\theta}_{NE}(\underline{x}))$, $BV(\hat{\theta}_{NE}(\underline{x}))$ and $\hat{V}_S(\hat{\theta}_{NE}(\underline{x}))$. Since $\hat{V}_S(\hat{\theta}_{NE}(\underline{x}))$ is close to the true Variance, $V(\hat{\theta}_{NE}(\underline{x}))$ [see Appendix (3.3)] the results in Table 3(vii) suggest that $BV(\hat{\theta}_{NE}(\underline{x}))$ is a better approximation to $V(\hat{\theta}_{NE}(\underline{x}))$.

(3.10) Summary and Further Discussions

The empirical distribution of $\hat{\theta}(\underline{x})$ was generated under varying situations and we estimated the first four moments. The bias of $\hat{\theta}(\underline{x})$ played a central role in determining the confidence probabilities

obtained. In particular, when $\Omega_1 \neq \Omega_2$, $\hat{\epsilon}_1$ could be very large and STAT1 using narrower intervals [see (3.2.1)] is less likely to capture the true log-odds, θ_T. STAT2 on the other hand always has the advantage of unbiased estimates of θ, but because of larger variances will have wider intervals. Also he tends to have confidence probabilities greater than the nominal value of 0.95 for small sample sizes and large p.

Changing the parameters $\underline{\mu}$, Δ and $D=\text{diag}(d_1,\ldots,d_p)$ has the crucial effect of changing the bias of $\hat{\theta}_E(\underline{x})$ and consequently the performance of STAT1 with respect to confidence probabilities. This set of SIMULATIONS showed that the power of the TEST (see (3.1.5)) could be poor with $\hat{P}_0$ as low as 0.47 in one SIMULATION. STAT3 in such a case will "behave" like STAT1 for half of the time, and like STAT2 for the other half. Since $\hat{\epsilon}_1$ can be large, STAT3's estimate of θ_T should also be biased, therefore affecting his confidence probability.

There is evidence of skewness and kurtosis in the distribution of $\hat{\theta}(\underline{x})$ particularly for smaller(n_1, n_2) and large p. Although we may suspect that non-normality of the distribution of $\hat{\theta}(\underline{x})$ has an effect on the validity of the use of.

$$\hat{\theta}(\underline{x}) \pm 1.96 \sqrt{[\text{Var}(\hat{\theta}(\underline{x}))]}$$

the confidence probabilities are "close" to the nominal values of 0.95 for the larger sample sizes. We therefore believe that the intervals (3.2.1) and (3.2.2) are useful approximations when 'correct' assumptions are made regarding Ω_1, Ω_2 and when sample sizes are reasonably large with respect to p.

In view of the importance of the bias of $\hat{\theta}_E(\underline{x})$ when $\Omega_1 \neq \Omega_2$, we obtained an approximation to this bias, which we called BIAS1 (see 2.5.6). The SIMULATION results show BIAS1 to be a reasonable approximation.

Particularly for small (n_1, n_2), STAT2 using (3.2.2) tends to have large confidence probabilities (i.e. > 0.95) possibly partly due to $AV(\hat{\theta}_{NE}(\underline{x}))$ being overestimated. We therefore obtained another approximation to the variance of $\hat{\theta}_{NE}(\underline{x})$, i.e. $BV(\hat{\theta}_{NE}(\underline{x}))$ in Section (2.4.C). $BV(\hat{\theta}_{NE}(\underline{x}))$ showed some improvement over $AV(\hat{\theta}_{NE}(\underline{x}))$, but note that only a few cases were considered. We did not use $BV(\hat{\theta}_{NE}(\underline{x}))$ in any of our SIMULATIONS (given in Table 3(I)) but only note the existence of an alternative to $AV(\hat{\theta}_{NE}(\underline{x}))$. $BV(\hat{\theta}_{NE}(\underline{x}))$ was derived after completion of the simulation study. Another possible reason for $\hat{CP}2$ being too large could be the non-normality of the distribution of $\hat{\theta}$.

The choice of $\underline{x}$-point clearly affected the results. There does not seem to be a clear-cut relationship between the $\underline{x}$-point and, say, $\hat{CP}$. Perhaps the absence of this clear-cut relationship is a fair representation of the difficulty of this problem. Possibly no simple method can be expected to work well for all possible $\underline{x}$'s unless the sample sizes are large.

Admittedly, we have reported only a fraction of the "statistics" calculated in the SIMULATIONS. Nevertheless, we have looked at the key factors involved in studying the distribution of $\hat{\theta}(\underline{x})$, and the use of a particular type of interval estimate for θ_T. The other information obtained seemed to be of secondary importance and is not reported here.

p = 2

	$\underline{x}^T$	$n_1=400=n_2$	$n_1=40=n_2$	$n_1=20=n_2$	$n_1=40,\ n_2=20$	$n_1=20,\ n_2=40$	θ_T
$d_1=4$ $d_2=4$	$\underline{\mu}$	0.16, 0.16	0.52, 0.54	0.79, 0.81	0.76, 0.81	0.57, 0.57	0.61
	origin	0.08, 0.08	0.26, 0.26	0.40, 0.38	0.34, 0.31	0.34, 0.35	-1.88
	\$CC	0.07, 0.08	0.25, 0.24	0.38, 0.35	0.32, 0.31	0.32, 0.30	-1.39
	\$9085(≠)	0.22, 0.22	0.73, 0.75	1.10, 1.20	0.97, 1.01	0.90, 0.97	-0.98
$d_1=0.25$ $d_2=0.25$	$\underline{\mu}$	0.16, 0.16	0.52, 0.54	0.79, 0.84	0.76, 0.80	0.57, 0.56	3.39
	origin	0.57, 0.57	1.93, 1.95	2.98, 3.06	1.95, 2.01	2.97, 3.01	-6.61
	\$CC	0.11, 0.11	0.37, 0.38	0.53, 0.59	0.46, 0.50	0.46, 0.49	1.39
	\$9090	0.24, 0.24	0.80, 0.82	1.20, 1.29	1.02, 1.09	1.02, 1.08	1.39

TABLE (3(ii)) In each 'cell' or simulation, the two numbers from left to right are;

(a) $\sqrt{\{AV(\hat{\theta}_{NE})\}}$ (b) $\sqrt{\{\hat{V}_s(\hat{\theta}_{NE})\}}$

(≠): \$9090 does not exist here.

p = 5

	$\underline{x}^T$	$n_1=40=n_2$	$n_1=20=n_2$	θ_T
$d_i=4$	$\underline{\mu}$	0.63, 0.63	1.02, 1.05	-1.47
$i=1,\ldots,5$	origin	0.43, 0.39	0.75, 0.58	-3.97
	\$CC	0.42, 0.38	0.74, 0.58	-3.47
	\$9980	2.11, 2.17	3.44, 3.69	0.43
$d_i=0.25$	$\underline{\mu}$	0.63, 0.63	1.02, 1.03	5.47
$i=1,\ldots,5$	origin	2.03, 2.07	3.34, 3.45	-4.53
	\$CC	0.49, 0.48	0.80, 0.77	3.47
	\$9090	1.65, 1.71	2.67, 2.79	3.47

TABLE (3(ii)) Contd.

	$\underline{x}$-point	θ_T $\hat{P}_O$	$\hat{CP1}$ $\hat{CP2}$	$\hat{\epsilon 1}$ $\hat{\epsilon 2}$	$\sqrt{V_s}(\hat{\theta}_E)$ $\sqrt{V_s}(\hat{\theta}_{NE})$
SA(1,1)	$\underline{\mu}$	0.61 1.00	0.25 0.95	-0.18 0.00	0.06 0.16
	origin	-1.89 1.00	0.00 0.95	-1.08 0.00	0.09 0.08
	$CC	-1.39 1.00	0.00 0.95	-1.12 0.00	0.06 0.08
	$9085	-0.98 1.00	0.00 0.95	1.14 0.00	0.21 0.22
	$9038	0.42 1.00	0.20 0.95	0.29 0.00	0.11 0.18
SA(3,1)	$\underline{\mu}$	3.39 1.00	0.78 0.95	0.18 0.00	0.25 0.16
	origin	-6.61 1.00	0.00 0.95	-3.41 0.00	0.21 0.57
	$9090	1.39 1.00	0.00 0.96	-1.38 0.00	0.28 0.24
	$\underline{x}$-point	<- as above ->			$\sqrt{V}(\hat{\theta}_E)$ $\sqrt{\hat{V}_s}(\hat{\theta}_{NE})$
SA(2,1)	origin	-2.00 0.05	0.95 0.95	0.00 0.00	0.14 0.16
	$CC	0.00 0.05	0.95 0.95	0.00 0.00	0.07 0.09
	$9090	0.00 0.05	0.95 0.95	0.00 0.00	0.20 0.24
	$9038	1.81 0.05	0.95 0.95	0.00 0.00	0.17 0.18

TABLE (3(iii)a)

	x-point	$\hat{\theta}_T$ / $\hat{P}_o$	$\hat{CP}1$ / $\hat{CP}2$	$\hat{\epsilon}1$ / $\hat{\epsilon}2$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$ / $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SA(1,2)	$\underline{\mu}$	0.61 1.00	0.98 0.95	-0.17 0.00	0.21 0.54
	origin	-1.89 1.00	0.10 0.96	-1.05 0.00	0.32 0.26
	$CC	-1.39 1.00	0.01 0.96	-1.08 0.00	0.19 0.24
	$9085	-0.98 1.00	0.51 0.98	1.19 0.00	0.69 0.75
	$9038	0.42 1.00	0.86 0.95	0.32 0.00	0.36 0.62
SA(3,2)	$\underline{\mu}$	3.39 1.00	0.86 0.95	0.12 0.00	0.84 0.54
	origin	-6.61 1.00	0.02 0.94	-3.40 0.00	0.68 1.95
	$9090	1.39 1.00	0.59 0.98	-1.44 0.00	0.91 0.82
	x-point	<- as above ->			$\sqrt{\hat{V}(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SA(2,2)	origin	-2.00 0.05	0.94 0.95	0.00 0.00	0.46 0.54
	$CC	0.00 0.05	0.96 0.96	0.00 0.00	0.23 0.29
	$9090	0.00 0.05	0.96 0.98	0.00 0.00	0.66 0.83
	$9038	1.81 0.05	0.95 0.95	0.00 0.00	0.54 0.62

TABLE (3(iii)b)

		θ_T	$\hat{CP1}$	$\hat{\epsilon 1}$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$
	x-point	$\hat{P}_o$	$\hat{CP2}$	$\hat{\epsilon 2}$	$\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SA(1,3)	μ	0.61 0.94	0.99 0.94	-0.14 0.00	0.31 0.81
	origin	-1.89 0.94	0.28 0.98	-1.00 0.00	0.47 0.38
	$CC	-1.39 0.94	0.10 0.98	-1.05 0.00	0.29 0.35
	$9085	-0.98 0.94	0.81 0.99	1.26 0.00	1.05 1.20
	$9038	0.42 0.94	0.91 0.94	0.36 0.00	0.54 0.95
SA(3,3)	μ	3.39 0.94	0.87 0.95	0.04 0.00	1.25 0.84
	origin	-6.61 0.94	0.15 0.93	-3.39 0.00	1.02 3.06
	$9090	1.39 0.94	0.84 1.00	-1.49 0.00	1.39 1.29
	x-point	<- as above ->			$\sqrt{V(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SA(2,3)	origin	-2.00 0.05	0.94 0.95	0.00 0.00	0.68 0.81
	$CC	0.00 0.05	0.98 0.98	0.00 0.00	0.34 0.43
	$9090	0.00 0.05	0.97 1.00	0.00 0.00	0.97 1.28
	$9038	1.81 0.05	0.95 0.95	0.00 0.00	0.79 0.98

TABLE (3(iii)C)

		θ_T	$\hat{CP1}$	$\hat{\epsilon 1}$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$
	x-point	$\hat{P}_O$	$\hat{CP2}$	$\hat{\epsilon}_2$	$\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SA(1,4)	μ	0.61 0.99	0.89 0.95	-0.35 0.00	0.31 0.57
	origin	-1.89 0.99	0.43 0.97	-0.79 0.00	0.56 0.35
	$CC	-1.39 0.99	0.16 0.97	-0.98 0.00	0.36 0.30
	$9085	-0.98 0.99	0.51 0.98	1.81 0.00	1.17 0.97
	$9038	0.42 0.99	0.89 0.96	0.32 0.00	0.60 0.66
SA(3,4)	μ	3.39 0.99	0.69 0.96	0.65 0.00	0.73 0.56
	origin	-6.61 0.99	0.01 0.93	-3.97 0.00	0.65 3.01
	$9090	1.39 0.99	0.84 0.98	-0.96 0.00	0.82 1.08
	x-point	<- as above ->			$\sqrt{\hat{V}(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SA(2,4)	origin	-2.00 0.05	0.94 0.93	0.00 0.00	0.61 0.79
	$CC	0.00 0.05	0.96 0.97	0.00 0.00	0.29 0.37
	$9090	0.00 0.05	0.96 0.98	0.00 0.00	0.80 1.08
	$9038	1.81 0.05	0.95 0.96	0.00 0.00	0.61 0.67

TABLE (3(iii)d)

	$\underline{x}$-point	θ_T $\hat{P}_O$	$\hat{CP1}$ $\hat{CP2}$	$\hat{\epsilon 1}$ $\hat{\epsilon 2}$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SA(1,5)	$\underline{\mu}$	0.61 0.99	0.99 0.94	-0.02 0.00	0.21 0.81
	origin	-1.89 0.99	0.05 0.97	-1.17 0.00	0.27 0.31
	$CC	-1.39 0.99	0.00 0.97	-1.12 0.00	0.17 0.31
	$9085	-0.98 0.99	0.79 0.97	0.84 0.00	0.62 1.01
	$9038	0.42 0.99	0.86 0.94	0.36 0.00	0.33 0.92
SA(3,5)	$\underline{\mu}$	3.39 0.99	0.89 0.93	-0.80 0.00	1.37 0.80
	origin	-6.61 0.99	0.26 0.94	-2.53 0.00	1.04 2.01
	$9090	1.39 0.99	0.57 0.97	-2.25 0.00	1.55 1.09
SA(2,5)	$\underline{x}$-point	<- as above ->			$\sqrt{\hat{V}(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
	origin	-2.00 0.05	0.94 0.96	0.00 0.00	0.51 0.56
	$CC	0.00 0.05	0.96 0.97	0.00 0.00	0.29 0.37
	$9090	0.00 0.05	0.96 0.98	0.00 0.00	0.80 1.08
	$9038	1.81 0.05	0.94 0.93	0.00 0.00	0.69 0.95

TABLE (3(iii)e)

	x-point	$\hat{\theta}_T$	$\hat{CP1}$	$\hat{\epsilon 1}$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$
		$\hat{P}_O$	$\hat{CP2}$	$\hat{\epsilon 2}$	$\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SE(1,1)	$\underline{\mu}$	-1.47 1.00	0.00 0.96	-2.21 0.00	0.22 0.63
	origin	-3.97 1.00	0.00 0.97	-3.06 0.00	0.33 0.39
	$CC	-3.47 1.00	0.00 0.97	-3.11 0.00	0.21 0.38
	$9980	0.43 1.00	0.01 0.96	3.41 0.00	1.03 2.17
SE(3,1)	$\underline{\mu}$	5.47 1.00	0.23 0.96	2.11 0.00	0.85 0.63
	origin	-4.53 1.00	0.47 0.94	-1.32 0.00	0.71 2.07
	$CC	3.47 1.00	0.02 0.97	2.29 0.00	0.44 0.48
	$9090	3.47 1.00	0.63 0.98	-2.33 0.00	1.56 1.71
	x-point	<- as above ->			$\sqrt{\hat{V}(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SE(2,1)	origin	-2.00 0.05	0.94 0.96	0.00 0.00	0.48 0.63
	$CC	0.00 0.05	0.97 0.97	0.00 0.00	0.24 0.41
	$9090	0.00 0.05	0.96 0.98	0.00 0.00	0.98 1.68

TABLE (3(iii)f)

		θ_T	$\hat{CP1}$	$\hat{\epsilon 1}$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$
	x-point	$\hat{P_o}$	$\hat{CP2}$	$\hat{\epsilon 2}$	$\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SE(1,2)	μ	-1.47 0.99	0.00 0.98	-2.15 0.00	0.36 1.05
	origin	-3.97 0.99	0.02 0.99	-2.95 0.00	0.52 0.58
	$CC	-3.47 0.99	0.00 0.99	-3.02 0.00	0.36 0.58
	$9980	0.43 0.99	0.16 0.98	3.59 0.00	1.61 3.69
SE(3,2)	μ	5.47 0.99	0.46 0.98	1.92 0.00	1.36 1.03
	origin	-4.53 0.99	0.63 0.93	-1.29 0.00	1.11 3.45
	$CC	3.47 0.99	0.14 0.99	2.17 0.00	0.71 0.77
	$9090	3.47 0.99	0.86 1.00	-2.53 0.00	2.40 2.79
	x-point	<- as above ->			$\sqrt{\hat{V}(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$
SE(2,2)	origin	-2.00 0.05	0.93 0.98	0.00 0.00	0.72 1.02
	$CC	0.00 0.05	0.98 0.99	0.00 0.00	0.37 0.64
	$9090	0.00 0.05	0.97 1.00	0.00 0.00	1.48 2.85

TABLE (3(iii)g)

	x-point	θ_T $\hat{P}_o$ -	$\hat{CP}1$ $\hat{CP}2$ $\hat{CP}3$	$\hat{\epsilon}1$ $\hat{\epsilon}2$ $\hat{\epsilon}3$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$ $\sqrt{\hat{V}_s(\hat{\theta}_3)}$
SB(2,1)	$\underline{\mu}$	1.31 0.80 -	0.35 0.93 0.86	0.66 0.00 0.04	0.22 0.81 0.82
	origin	-1.19 0.80 -	0.39 0.97 0.88	-0.51 0.00 -0.05	0.27 0.31 0.35
	$CC	-0.69 0.80 -	0.24 0.97 0.83	-0.45 0.00 -0.06	0.17 0.30 0.34
	$9090	-0.69 0.80 -	0.91 0.97 0.96	-0.45 0.00 -0.08	0.68 1.09 1.08
	$9038	1.11 0.80 -	0.70 0.93 0.91	0.53 0.00 0.02	0.38 0.94 0.93
SB(3,1)	$\underline{\mu}$	2.00 0.99 -	0.03 0.93 -	1.32 0.00 -	0.22 0.82 -
	origin	-0.50 0.99 -	0.93 0.97 -	0.15 0.00 -	0.27 0.31 -
	$CC	0.00 0.99 -	0.78 0.97 -	0.21 0.00 -	0.18 0.30 -
	$9090	0.00 0.99 -	0.87 0.98 -	-0.58 0.00 -	0.60 1.08 -
	$9038	1.81 0.99 -	0.22 0.93 -	1.07 0.00 -	0.35 0.92 -

TABLE (3(iv)a)

	x-point	$\hat{\theta}_T$ $\hat{P}_O$ -	$\hat{CP}1$ $\hat{CP}2$ $\hat{CP}3$	$\hat{\epsilon}1$ $\hat{\epsilon}2$ $\hat{\epsilon}3$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$ $\sqrt{\hat{V}_s(\hat{\theta}_3)}$
SB(3,2)	μ	2.69 0.84 -	0.70 0.93 0.91	0.66 0.00 0.03	0.61 0.83 0.85
	origin	-1.31 0.84 -	0.85 0.96 0.94	0.65 0.00 0.06	0.50 0.55 0.60
	$CC	0.69 0.84 -	0.35 0.97 0.89	0.66 0.00 0.05	0.30 0.36 0.43
	$9090	0.69 0.84 -	0.81 0.98 0.95	-1.06 0.00 -0.11	0.89 1.07 1.14
	$9038	2.50 0.84 -	0.87 0.93 0.93	0.28 0.00 -0.02	0.73 0.95 0.94
SD(1,1)	origin	-18.42 0.05 -	0.94 0.93 -	0.00 0.00 -	4.81 7.07 -
	$CC	0.00 0.05 -	0.97 1.00 -	0.00 0.00 -	1.02 2.50 -
	$x_1 = 1.9037$ $x_2 = 0.9903$	-6.87 0.05 -	0.94 0.94 -	0.00 0.00 -	2.24 3.59 -
	The points $9090 and $9038 do not exist for SD(1,1)				

TABLE (3(iv)b)

	x-point	θ_T $\hat{P}_O$ –	$\hat{CP}1$ $\hat{CP}2$ $\hat{CP}3$	$\hat{\epsilon}1$ $\hat{\epsilon}2$ $\hat{\epsilon}3$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$ $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$ $\sqrt{\hat{V}_s(\hat{\theta}_3)}$
SC(1,1)	μ	2.00 0.99 –	0.91 0.93 –	-0.38 0.00 –	0.85 0.82 –
	origin	-4.25 0.99 –	0.15 0.94 –	-1.90 0.00 –	0.59 1.07 –
	$9090	0.00 0.99 –	0.93 0.98 –	0.34 0.00 –	0.43 1.07 –
	$9038	1.81 0.99 –	0.67 0.93 –	0.57 0.00 –	0.55 0.96 –
SC(1,2)	μ	2.00 0.99 –	0.51 0.93 –	-2.10 0.00 –	1.35 0.80 –
	origin	-8.00 0.99 –	0.08 0.94 –	-3.92 0.00 –	1.03 1.96 –
	$9090	0.00 0.99 –	0.96 0.98 –	-0.16 0.00 –	0.69 1.10 –
	$9038	1.81 0.99 –	0.87 0.93 –	-0.82 0.00 –	1.06 0.92 –

TABLE (3(iv)c)

	x-point	$\hat{\theta}_T$ / $\hat{P}_O$ / –	$\hat{CP}1$ / $\hat{CP}2$ / $\hat{CP}3$	$\hat{\epsilon}1$ / $\hat{\epsilon}2$ / $\hat{\epsilon}3$	$\sqrt{\hat{V}_s(\hat{\theta}_E)}$ / $\sqrt{\hat{V}_s(\hat{\theta}_{NE})}$ / $\sqrt{\hat{V}_s(\hat{\theta}_3)}$
SF(1,1)	μ	6.18 1.00 –	0.00 0.97 –	5.14 0.00 –	0.26 0.62 –
	origin	3.52 1.00 –	0.00 0.98 –	4.45 0.00 –	0.34 0.40 –
	$CC	4.18 1.00 –	0.00 0.97 –	4.39 0.00 –	0.21 0.38 –
	$9090	4.18 1.00 –	0.02 0.98 –	3.60 0.00 –	0.82 1.70 –
SF(1,2)	μ	2.00 0.47 –	0.47 0.96 0.71	0.67 0.00 0.28	0.34 0.64 0.66
	origin	-1.00 0.47 –	0.90 0.97 0.92	0.33 0.00 0.16	0.39 0.45 0.47
	$CC	0.00 0.47 –	0.84 0.97 0.89	0.23 0.00 0.10	0.21 0.39 0.36
	$9090	0.00 0.47 –	0.63 0.98 0.79	1.55 0.00 0.67	0.96 1.72 1.73

TABLE (3(iv)d)

	x-point	θ_T $\hat{P}_O$ -	$\hat{CP}1$ $\hat{CP}2$ $\hat{CP}3$	$\hat{\epsilon}1$ $\hat{\epsilon}2$ $\hat{\epsilon}3$	$\sqrt{\hat{V}_S(\hat{\theta}_E)}$ $\sqrt{\hat{V}_S(\hat{\theta}_{NE})}$ $\sqrt{\hat{V}_S(\hat{\theta}_3)}$
SG(1,1)	$\underline{\mu}$	2.00 0.46 -	0.94 0.96 0.95	0.00 0.00 -0.03	0.48 0.62 0.58
	origin	-2.23 0.46 -	0.87 0.96 0.92	-0.23 0.00 -0.07	0.48 0.68 0.62
	$CC	-2.03 0.46 -	0.86 0.96 0.91	-0.26 0.00 -0.10	0.43 0.65 0.59
	$9090	-1.73 0.46 -	0.85 0.97 0.90	1.28 0.00 0.59	1.12 2.02 1.86

TABLE (3(iv)e)

	$\underline{x}$-point	θ_T	$\hat{CP}1$	$\hat{\beta}1$	$\hat{\gamma}1$
		-	$\hat{CP}2$	$\hat{\beta}2$	$\hat{\gamma}2$
SA(2,3)	origin	-2.00	0.94	-0.95	4.55
		-	0.95	-1.44	7.33
	$CC	0.00	0.98	0.01	4.18
		-	0.98	0.07	4.21
	$9090	0.00	0.97	0.00	4.15
		-	1.00	0.04	7.64
	$9038	1.81	0.95	0.98	5.56
		-	0.95	1.76	10.31
SA(2,2)	origin	-2.00	0.94	-0.71	3.95
		-	0.95	-0.94	4.93
	$CC	0.00	0.96	0.01	3.68
		-	0.96	-0.02	3.38
	$9090	0.00	0.96	0.03	3.46
		-	0.98	-0.02	4.26
	$9038	1.81	0.95	0.56	3.77
		-	0.95	0.94	5.14
SE(2,2)	origin	-2.00	0.93	-1.08	5.43
		-	0.98	-1.52	11.10
	$CC	0.00	0.98	0.02	4.89
		-	0.99	-0.24	6.52
	$9090	0.00	0.97	-0.01	4.17
		-	1.00	0.02	6.26
SE(2,1)	origin	-2.00	0.94	-0.73	4.07
		-	0.96	-0.73	4.51
	$CC	0.00	0.97	0.06	3.53
		-	0.97	-0.01	3.02
	$9090	0.00	0.96	-0.02	3.34
		-	0.98	-0.01	4.09

TABLE (3(v)a)

x-point	$\hat{CP}2$	$\hat{\beta}2$	$\hat{\gamma}2$	θ_T	Simulation
$\underline{\mu}$	0.94	1.53	7.83	0.61	SA(1,3) $p = 2$ $n_1 = 20 = n_2$ $d_i = 4$ $(i = 1,2)$
origin	0.98	-0.50	4.54	-1.88	
$CC	0.98	0.01	3.22	-1.39	
$9085	0.99	0.01	9.38	-0.98	
$9038	0.94	1.59	9.20	0.42	
$\underline{\mu}$	0.95	0.93	4.85	0.61	SA(1,2) $p = 2$ $n_1 = 40 = n_2$ $d_i = 4$ $(i = 1,2)$
origin	0.96	-0.22	3.42	-1.88	
$CC	0.96	0.05	3.01	-1.39	
$9085	0.98	0.20	4.07	-0.98	
$9038	0.95	0.94	5.26	0.42	
$\underline{\mu}$	0.98	1.96	15.97	-1.47	SE(1,2) $p = 5$ $n_1 = 20 = n_2$ $d_i = 4$ $(i = 1,..,5)$
origin	0.99	-0.24	3.64	-3.97	
$CC	0.99	0.02	3.27	-3.47	
$9980	0.98	1.60	11.95	0.43	
$\underline{\mu}$	0.96	0.81	5.20	-1.47	SE(1,1) $p = 5$ $n_1 = 40 = n_2$ $d_i = 4$ $(i = 1,..,5)$
origin	0.97	-0.05	3.09	-3.97	
$CC	0.97	0.01	3.07	-3.47	
$9980	0.96	0.72	4.96	0.43	

TABLE (3(v)b) Effect of skewness and Kurtosis on $\hat{CP}2$

	x-point	θ_T	$\hat{CP}1$	$\hat{\beta}1$	$\hat{\gamma}1$
		-	$\hat{CP}2$	$\hat{\beta}2$	$\hat{\gamma}2$
SA(1,1)	μ	0.61	0.25	0.22	3.13
		-	0.95	0.27	3.24
	origin	-1.89	0.00	-0.28	3.21
		-	0.95	0.00	2.99
	$CC	-1.39	0.00	-0.27	3.11
		-	0.95	0.02	2.94
	$9085	-0.98	0.00	-0.21	3.13
		-	0.95	0.05	3.03
	$9038	0.42	0.20	-0.10	3.04
		-	0.95	0.22	3.07
SA(2,1)	origin	-2.00	0.95	-0.20	3.01
		-	0.95	-0.24	3.10
	$CC	0.00	0.95	0.01	3.00
		-	0.95	-0.04	2.93
	$9090	0.00	0.95	0.00	3.07
		-	0.95	0.01	3.11
	$9038	1.81	0.95	0.16	3.07
		-	0.95	0.24	3.08

TABLE (3(v)c)

Simulation / $\underline{x}$-Point	SA(1,1)	SA(1,2)	SA(1,3)
$\underline{\mu}$	-0.19(-0.18)	-0.19(-0.17)	-0.20(-0.14)
origin	-1.08(-1.08)	-1.05(-1.05)	-1.01(-1.00)
$CC	-1.12(-1.12)	-1.09(-1.08)	-1.06(-1.05)
$9085	1.14 (1.14)	1.16 (1.19)	1.17 (1.26)
$9038	0.29 (0.30)	0.29 (0.32)	0.29 (0.36)
	SA(3,1)	SA(3,2)	SA(3,3)
μ	0.17 (0.18)	0.01 (0.12)	-0.15 (0.04)
origin	-3.41(-3.41)	-3.41(-3.40)	-3.42(-3.39)
$CC	0.31 (0.32)	0.23 (0.27)	0.14 (0.22)
$9090	-1.40(-1.38)	-1.56(-1.44)	-1.74(-1.49)
$9038	-0.31(-0.30)	-0.49(-0.37)	-0.67(-0.48)
	SE(1,1)	SE(1,2)	
$\underline{\mu}$	-2.27(-2.21)	-2.26(-2.15)	
origin	-3.08(-3.06)	-3.00(-2.95)	
$CC	-3.13(-3.11)	-3.06(-3.02)	
$9980	3.13 (3.41)	3.00 (3.59)	

TABLE (3(vi)): Comparing BIAS1 with $\hat{\epsilon}1$, the latter in Parenthesis.

$\underline{x}$-point	$\sqrt{AV(\hat{\theta}_{NE}(\underline{x}))}$	$\sqrt{BV(\hat{\theta}_{NE}(\underline{x}))}$	$\sqrt{\hat{V}_s(\hat{\theta}_{NE}(\underline{x}))}$	Parameters
$\underline{\mu}$	0.524	0.538	0.536	$n_1=40=n_2$
origin	0.264	0.259	0.259	$p=2$
\$CC	0.248	0.242	0.240	$\Omega_1=4\Omega_2$
\$9038	0.598	0.618	0.621	$\Omega_2\equiv I_2$
\$9085	0.735	0.762	0.748	$\underline{\mu}^T=(2,0)$
$\underline{\mu}$	0.788	0.834	0.815	as above
origin	0.402	0.385	0.384	except
\$CC	0.378	0.355	0.351	$n_1=20=n_2$
\$9038	0.895	0.961	0.945	
\$9085	1.099	1.189	1.203	

Table (3(vii))

Note: $AV\{\hat{\theta}_{NE}(\underline{x})\}$, $BV\{\hat{\theta}_{NE}(\underline{x})\}$ and $\hat{V}_s\{\hat{\theta}_{NE}(\underline{x})\}$ are defined in (2.3.1), section (2.4.C), (3.2.4) respectively.

CHAPTER 4

Analysis of data sets: using approximate interval estimate techniques

4.1 Introduction

The central theme of our work has been to study the distribution of the log-odds ratio, $\hat{\theta}(\underline{x})$, and to construct interval estimates for $\theta(\underline{x})$. The confidence interval $[\hat{\theta}(\underline{x}) \pm 1.96 \sqrt{var(\hat{\theta}(\underline{x})}]$ was used to express our uncertainty when estimating $\theta(\underline{x})$. We now try out our ideas on specific data sets.

For each data set we will report the result of the likelihood ratio test for testing the equality of the variance-covariance matrices (see Appendix 4.3). We will use (n_1, n_2) to denote the sample sizes from the training sets. The test set will have size n_3. The dimensionality parameter will be p (i.e dimension of $\underline{x}$).

DATA SET 1 IRIS DATA ($n_1 = 50$, $n_2 = 50$, $p = 4$)

Fisher (1936) used this data set to study the linear discriminant function. Fifty random observations on each of three species of a flower, IRIS SETOSA, IRIS VERSICOLOUR, and IRIS VIRGINICA were studied. Based on four measurements, sepal length, sepal width, petal length and petal width, we would like to know the species of a particular flower. For our purposes, we will concentrate only on the two species that have a slight 'overlap', IRIS VERSICOLOUR and IRIS VIRGINICA. The actual data set was obtained from Dixon, (1977), page 712, and is reproduced here in Appendix 4.2.

The likelihood ratio test is significant, the test statistic is 35.04 against the critical value of 18.31 ($= \chi^2(10; 0.95)$).

DATA SET 2 HAEMOPHILIA ($n_1 = 20$, $n_2 = 23$, $n_3 = 7$, $p = 2$)

Carriers of haemophilia suffer from a bleeding disorder due to

deficiency of clotting agents in the blood. Haemophilia is a life-long, crippling disorder, and carriers are informed of the consequences of their having children. As the treatment of haemophiliacs places a great strain on the blood transfusion services, genetic counselling for haemophiliacs is obviously important. Identification of haemophilia carriers is based on two measurements; (i) factor VIII-related antigen and (ii) factor VIII activity. The measurements were obtained from 20 carriers and 23 normal women (controls). There are seven unclassified patients ($n_3 = 7$).

The data set was obtained from Prentice, C.R.M. et al (1975), and is reproduced in Appendix (4.2.c).

The likelihood ratio test is significant, with the test statistic equal to 9.73 against a critical value of 7.82 (χ^2 (3;0·95))

DATA SET 3 CONNS DATA ($n_1 = 20$, $n_2 = 11$, $n_3 = 4$, $p = 4$)

A rare disease called CONN'S Syndrome can be due to

(i) a benign tumour (adenoma) in the adrenal cortex, or

(ii) a more diffuse condition (bilateral hyperplasia) of the adrenal glands.

The treatment may involve either, the removal of an adenomatous adrenal gland or drug therapy. It is therefore necessary to know which group a patient belongs to; i.e. either adenoma or bilateral hyperplasia. Assessments of patient-type in other studies (Aitchison, Habbema, Kay (1977)) were based on eight measurements on the patients. For our study we consider only four; AGE, POTASSIUM, CARBON DIOXIDE and RENIN. We note that our variance formulae requires $n_i > p+4$. The logarithms (base e) of the data were used. Thirty-one patients who were operated on gave confirmation on twenty cases of adenoma ($n_1 = 20$) and eleven

cases of bilateral hyperplasia (n_2 = 11). We have four patients in our test set (n_3 = 4).

The data set was obtained from Aitchison, and Dunsmore, (1975), page 10, and is reproduced here in Appendix (4.2.B).

The likelihood ratio test is just non-significant, with the test statistic equal to 17.92 against the critical value of 18.31 (= χ^2 (10; 0.95)). Further investigation into the marginal distributions of the data showed that one of the four variables had significant differences for the sample variances. We regard the likelihood ratio test with reservation, bearing in mind that for n_1 = 20, n_2 = 11 and p = 4, the large sample distribution of the test statistic cannot be justified (Box (1949)).

4.2 Graphical procedures for $\Omega_1 = \Omega_2$ case

(A) Introduction

Firstly, we note that most of the variables used here have been defined in Chapter 2. Using the canonical form, as illustrated in (3.1.1), Critchley and Ford (1985) showed that the sampling distribution of $\hat{e}_E(\underline{x})$ can be defined in terms of Δ and $\underline{x}^*$. Using symmetry about the x_1^*-axis they also showed that the distribution of $\hat{e}_E(\underline{x})$ is invariant under rotations about that axis. The distribution of $\hat{e}_E(\underline{x})$ can then be parameterised in terms of,

$$(\Delta,\ x_1^*,\ \underline{x}^{*T}\underline{x}^*) \text{ or equivalently by } (\Delta,\ \alpha_1(\underline{x}),\ \alpha_2(\underline{x})).$$

Critchley and Ford (1985) showed that $e(\underline{x}) = \Delta x_1^*$, where x_1^* is the first component of $\underline{x}^*$. The parameter space $(\Delta, x_1^*, \underline{x}^{*T}\underline{x}^*)$ is for a given Δ equivalent to (x_1^*, y_1^*), where $y_1^* = (\underline{x}^{*T}\underline{x}^* - x_1^{*2})^{1/2}$. They in fact considered only the half-plane $(x_1^*, y_1^* | y_1^* \geq 0)$. In appendix (4.1), using the equivalent parameter space $(\Delta, \alpha_1(\underline{x}), \alpha_2(\underline{x}))$ it can be explicitly shown that,

$$y_1^* = \left[\Phi(\underline{x}) - \frac{\Delta^2}{4} - \frac{\theta^2}{\Delta^2}\right]^{1/2}$$

and as stated above

$$x_1^* = \theta(\underline{x})/\Delta.$$

When analysing our data set, we of course need to estimate x_1^* and y_1^*. We estimate Δ, θ and Φ using data from the two training sets. Having obtained $\hat{\Delta}$, $\hat{\theta}$, $\hat{\Phi}$ we can then calculate $\hat{x}_1^*$, $\hat{y}_1^*$ corresponding to an $\underline{x}$ point from the test set. It is of interest to plot points from the training set together with those from the test-set.

The 'training' points can be used to check on.

(i) misclassification properties

(ii) distributional assumptions; e.g. equality of covariance matrices

(iii) the "GREY AREA", see figure 4(i).

Henceforth any reference to two training samples of sizes n_1 and n_2 is equivalent to having n_1 plus n_2 values of $\hat{x}_1^*$ and $\hat{y}_1^*$.

For purposes of notation, members of group 1 and group 2, i.e. the training samples, will be labelled with the numbers 1 and 2 respectively. Any observation from the test-set (unclassified point) will be labelled with the letters A, B, C,...., etc.

(B) Critchley/Ford plot

For the equal covariance case, Critchley and Ford (1985) have an informative plot, viz:-

$(x_1^*, y_1^* | y_1^* \geq 0)$

From section (4.2.A) above we see that

(i) x_1^* is proportional to the discriminant score, i.e. $\theta(\underline{x})$

(ii) for fixed x_1^*, $V(\hat{\theta}_E)$ is increasing in y_1^*.

We obtain further information from the (x_1^*, y_1^*) plot by including tolerance regions [Guttman (1970)] for each population.

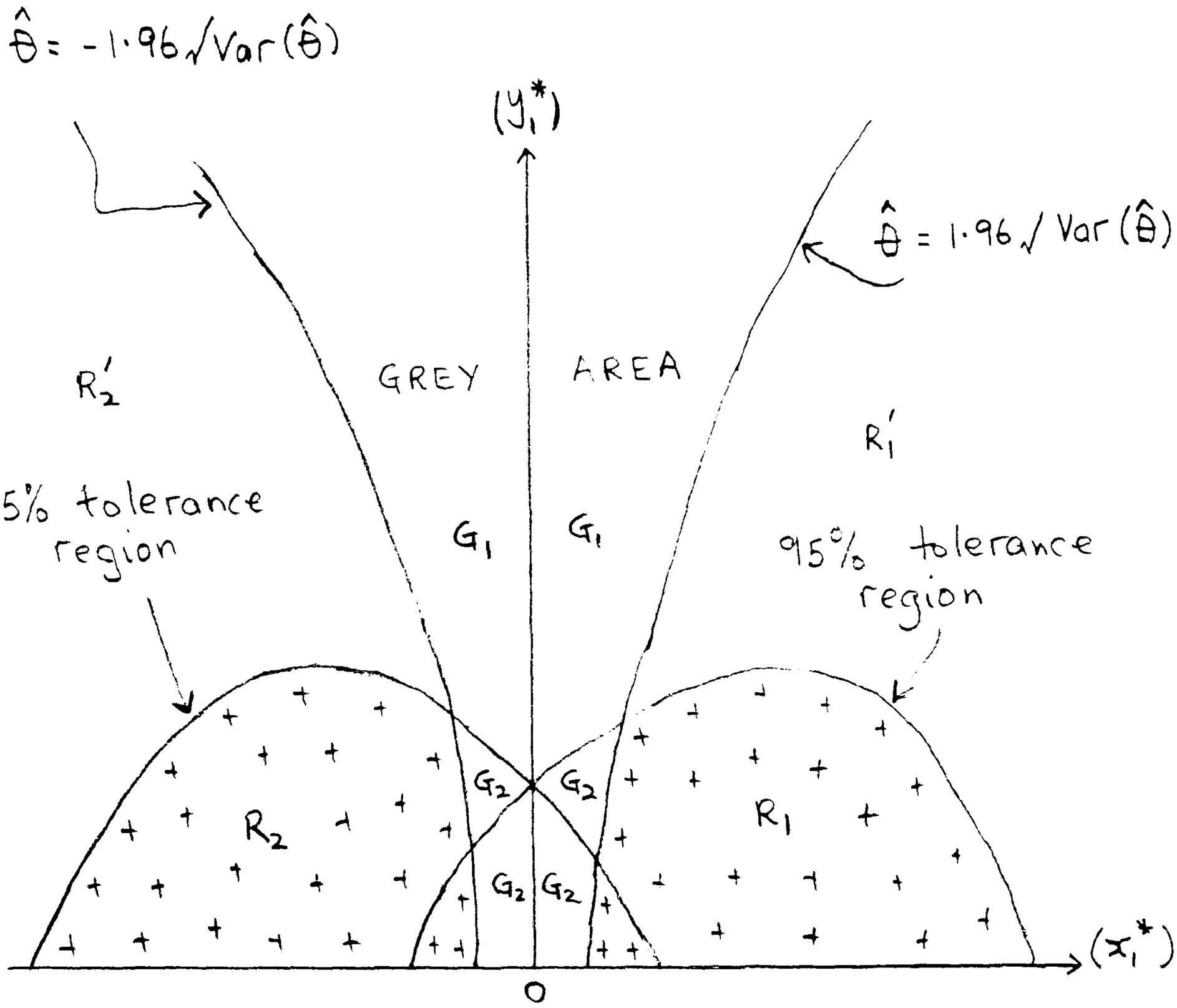


Figure (4(i)): The Critchley/Ford Plot

We use the following notation to describe the various regions of the diagram above.

(i) R_1 and R_2: as shown shaded with '+' signs. Interval estimate does not contain zero and x-points not atypical.

(ii) R_1' and R_2': points are atypical but Interval estimate does not contain zero.

(iii) G_1: inside GREY AREA but outside tolerance regions. Points here are atypical.

(iv) G_2: interval estimate contains zero, points not atypical.

The tolerance regions are defined as follows.

$$\hat{\alpha}_i^2(\underline{x}) = F(p,\ n_1+n_2 - p-1;\ q).\frac{p(n_1+n_2)(1+\frac{1}{n_i})}{(n_1+n_2 - p-1)}$$

$(i = 1,2)$

where $\hat{\alpha}_i^2(\underline{x}) = (\underline{x} - \underline{\bar{x}}_i)^T\ [(n_1+n_2)(S_1+S_2)^{-1}]\ (\underline{x} - \underline{\bar{x}}_i)$ and $F(p,\ n_1+n_2 - p-1;\ q)$ is the q x 100 percentile of the $F(p,\ n_1+n_2 - p-1)$ distribution. In our plots we shall consider the 95% and 99% tolerance regions.

Further, the confidence interval $(\hat{\theta} \pm 1.96\ \sqrt{var(\hat{\theta})})$ will contain the value zero if and only if

$$\left|\frac{\hat{\theta}}{\sqrt{var(\hat{\theta})}}\right| < 1.96.$$

In the hypothetical plot, figure 4(I), a value of $\hat{\theta}(\underline{x})$ is significantly non-zero if it lies outside the "grey area", i.e. below the contours where $\frac{\hat{\theta}}{\sqrt{var(\hat{\theta})}}$ takes the values $\pm$ 1.96.

Figure 4(I) gives a brief description of various regions in the $(x_1^*,\ y_1^*|y_1^* \geq 0)$ half-plane. We will count the number of $\underline{x}$-points, from each data set, that fall into these regions, and tabulate their totals.

(C) Alternative plot

In Section (4.2.B) we stated that

(i) x_1^* is proportional to $\theta(\underline{x})$

(ii) Variance of $\hat{\theta}$ is a function of y_1^*.

Thus an alternative to the $(x_1^*,\ y_1^*)$ plot is simply to plot $\sqrt{V(\hat{\theta})}$ against $\hat{\theta}$. We shall call this plot the $(\hat{\theta},\ SD(\hat{\theta}))$ plot where $SD(\hat{\theta}) = \sqrt{V(\hat{\theta})}$. The subscript 'E' or 'NE' of $\hat{\theta}_{(.)}$ will denote the assumption of equality or inequality of population covariance matrices.

We will not be able to include the tolerance regions in the $[\hat{\theta},\ SD(\hat{\theta})]$ plot. We will however keep the contours $\hat{\theta} = \pm$ 1.96 $SD(\hat{\theta})$.

The equivalent of figure 4(I) for the $(\hat{\theta}, SD(\hat{\theta}))$ plot reduces the half-plane to three regions.

Firstly R1: interval completely positive

next, R2: interval completely negative

and finally, GREY AREA: interval contains zero.

When analysing data sets, Tables 4(I), 4(II), ..., 4(v) give the count of $\underline{x}$-points that fall into the various categories.

(D) Applications to examples

	R_1	R_1'	R_2	R_2'	G_1	G_2	
Type 1	46	0	1*	0	0	3	$n_1 = 50$
Type 2	0*	0	44	3	0	3	$n_2 = 50$

Table 4(i) IRIS DATA [95% tolerance region]

'Cell'-counts from figure 4(ii).

(* = number misclassified)

The data for type 2 is slightly more variable than that for type 1, thus 3 type 2 observations lie outside the 95% tolerance region. We would be uncertain in allocating six observations: three from each type ($\equiv G_2$). There is one type 1 observation misclassified. Generally most of the $\underline{x}$ - points are correctly assigned.

	R_1	R_1'	R_2	R_2'	G_1	G_2	
Type 1	12	0	1	0	0	7	$n_1 = 20$
Type 2	0	0	8	0	0	3	$n_2 = 11$

Table 4(ii) CONNS DATA [95% tolerance region]

'Cell'-counts from figure 4(iii)

We have far more observations that we are uncertain about. Only one type 1 is misclassified.

Generally the discrimination power is less than for the previous data set.

	R_1	R_2	Grey Area
Type 1	46	1	3
Type 2	0	47	3

Table 4(iii)a: IRIS DATA

Cell counts from figure 4(iv)a

	R_1	R_2	Grey Area
Type 1	12	1	7
Type 2	0	8	3

Table 4(iii)b: CONNS DATA

Cell counts from figure 4(iv)b

We note that the (x_1^*, y_1^*) plots were used to show the various regions of uncertainty in allocating $\underline{x}$. The $[\hat{\theta}, SD(\hat{\theta})]$ plot was also included in this section only for comparison with the corresponding plots in the $\Omega_1 \neq \Omega_2$ case.

4.3: METHODS FOR $\Omega_1 \neq \Omega_2$ CASE

(A) Plot

When the population covariance matrices are not equal we do not have a simple two dimensional plot: in particular we do not

have the corresponding (x_1^*, y_1^*) plot. However, we can still proceed with the $(\hat{e}(\underset{\sim}{x}), SD(\hat{e}(\underset{\sim}{x})))$ plot.

(B) Analysis of Data

Since the haemophilia data is a two dimensional case, we firstly have a scatter plot of the raw data, in figure 4(v). Also included are the 7 test-set observations: A, B, ..., G. From the scatter plot we could suggest that

A - is a carrier

B - a borderline carrier

C - will very likely fall inside Grey Area

D - a possible outlier to both groups

E, F - borderline controls

G - borderline, but more probable a control

The $(\hat{e}, SD(\hat{e}))$ plot (with $\Omega_1 \neq \Omega_2$) gives us:

	R_1	R_2	Grey Area
Type 1	16	1	3
Type 2	0	18	5

TYPE 1 = carrier
TYPE 2 = control

$n_1 = 20$

$n_2 = 23$

Table 4(iv): HAEMOPHILIA DATA (excluding A, B, ..., G)

Cell counts from figure 4(vi)

One carrier is misclassified. Three carriers and 5 controls in Grey area. The presence of a possible outlier (TYPE 1) near point E[figure 4(v)] may account for 'inflated' variances which in turn, may have resulted in some of these 8 observations being inside the GREY AREA.

For the 7 test set observations, we have almost complete

agreement with the subjective impression from the scatter plot except;

E: falls just inside Grey Area

F: falls just outside Grey Area

G: a control

We note that type 2 members generally have small log-odds. It is perhaps wiser then not to make definite 'conclusions' for E and F. We come back to the CONN's data and do the $[\hat{\theta},\ SD(\hat{\theta})]$ plot (given $\Omega_1 \neq \Omega_2$).

	R_1	R_2	Grey Area	
Type 1	0	1	19	$n_1 = 20$
Type 2	0	9	2	$n_2 = 11$

Table 4(v): CONNS DATA

Cell counts from Figure 4(vii)

The most striking feature of figure 4(vii) is that none of the type 1 patients can be classified. Also note that patient A is now a borderline case $[-5.42 < \hat{\theta} < -0.001]$, where previously when we assume $\Omega_1 = \Omega_2$. A was very probably a group 2 member.

4.4 Summary remarks

Analysis of the CONN's data has shown that different assumptions made on the parameters of the distribution of $\underline{x}$ can substantially affect the subsequent discriminant analysis.

Reasonably large sample sizes (training set) relative to dimension (= p), e.g. for the IRIS and HAEMOPHILIA data, should ensure that the discrimination problem is 'safely' analysed. This

brings us back to the CONN's data. From figure 4(vii), it is 'felt' that there is no substantial overlap between the two groups and that it should be possible to classify some of the group 1's with some certainty. With $n_1 = 20$, $n_2 = 11$, $p = 4$, we would expect $\mathrm{Var}(\hat{e}(\underline{x}))$ to be poorly estimated. This is particularly true when we assume $\Omega_1 \neq \Omega_2$.

Given this concern aroused by this example (CONN's data), we now consider alternative procedures of interval estimation, which might be more reliable for small samples.

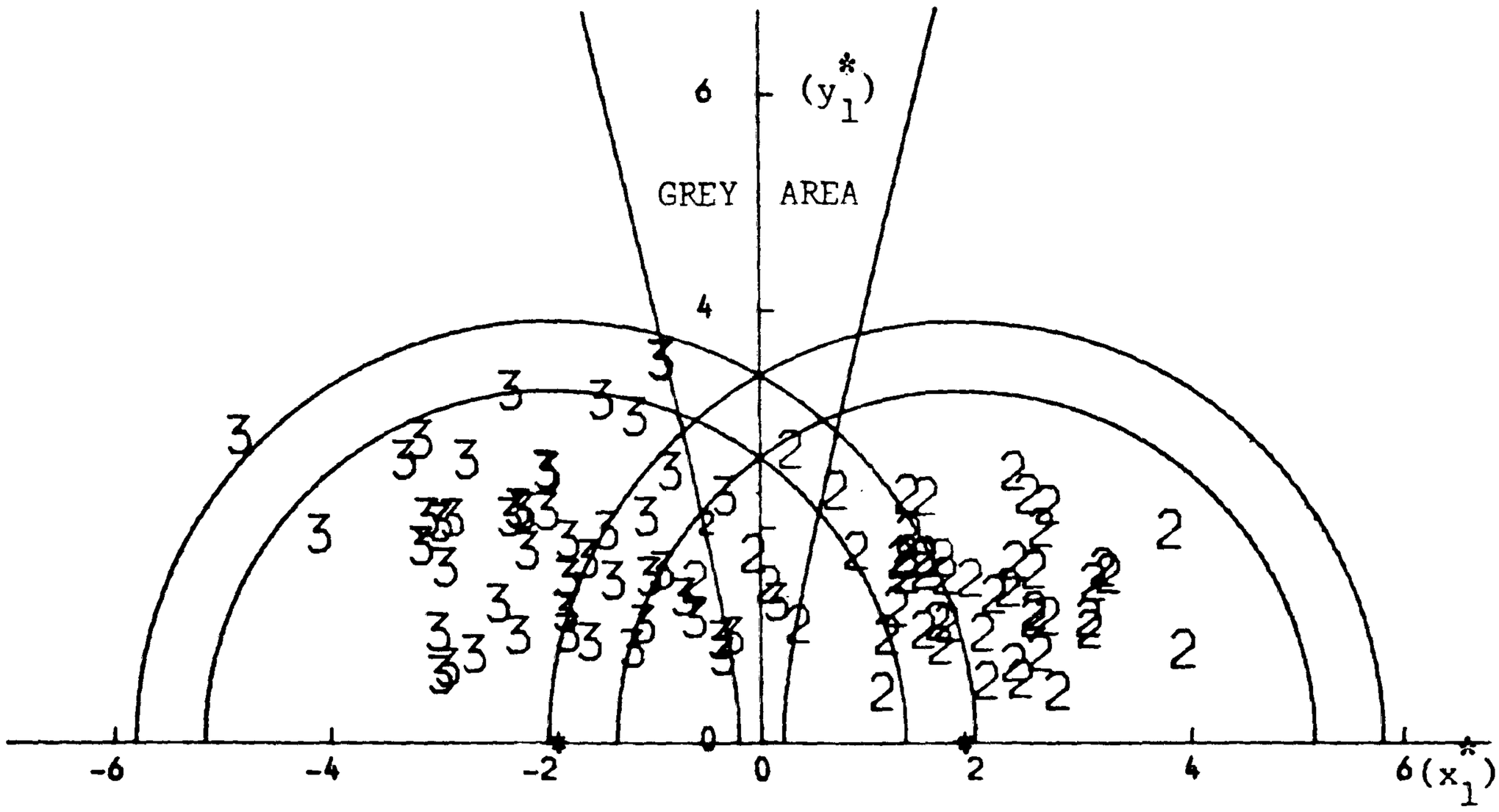


Figure (4(ii)): IRIS DATA

2 ≡ Iris Versicolour

3 ≡ Iris Virginica

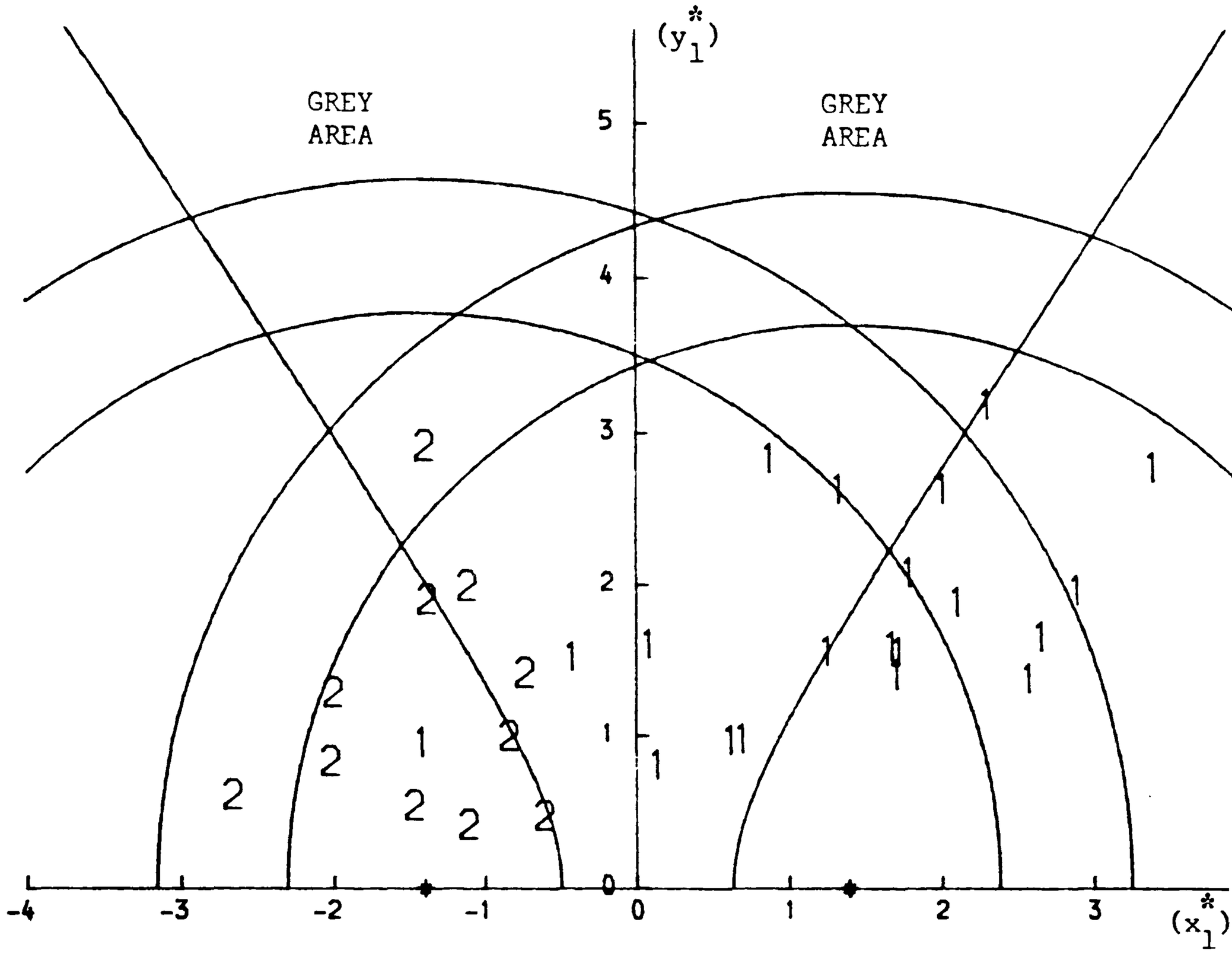


Figure (4(iii)): CONN'S DATA

1 ≡ Adenoma

2 ≡ Bilateral hyperplasia

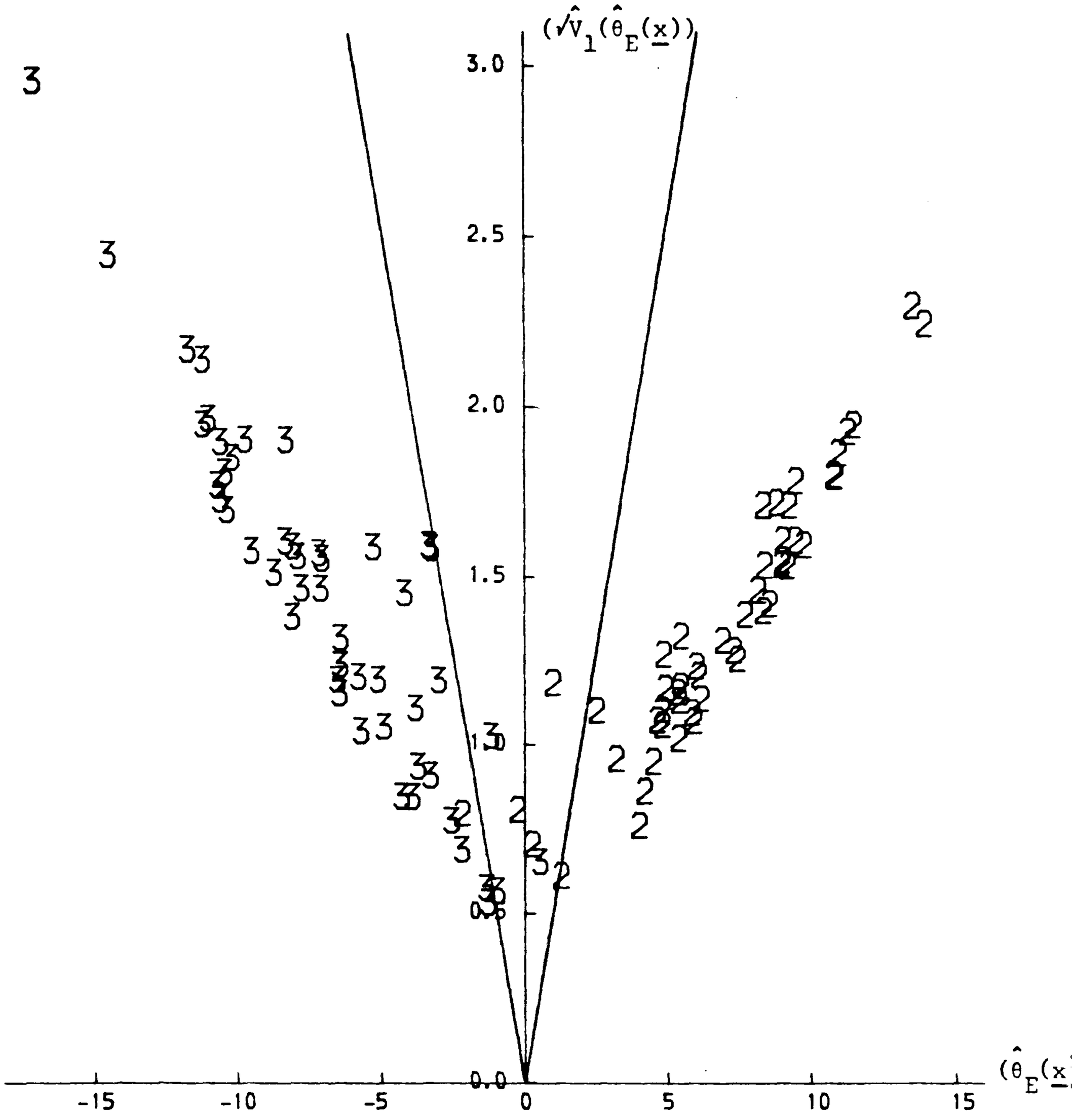


Figure (4(iv)a): IRIS DATA

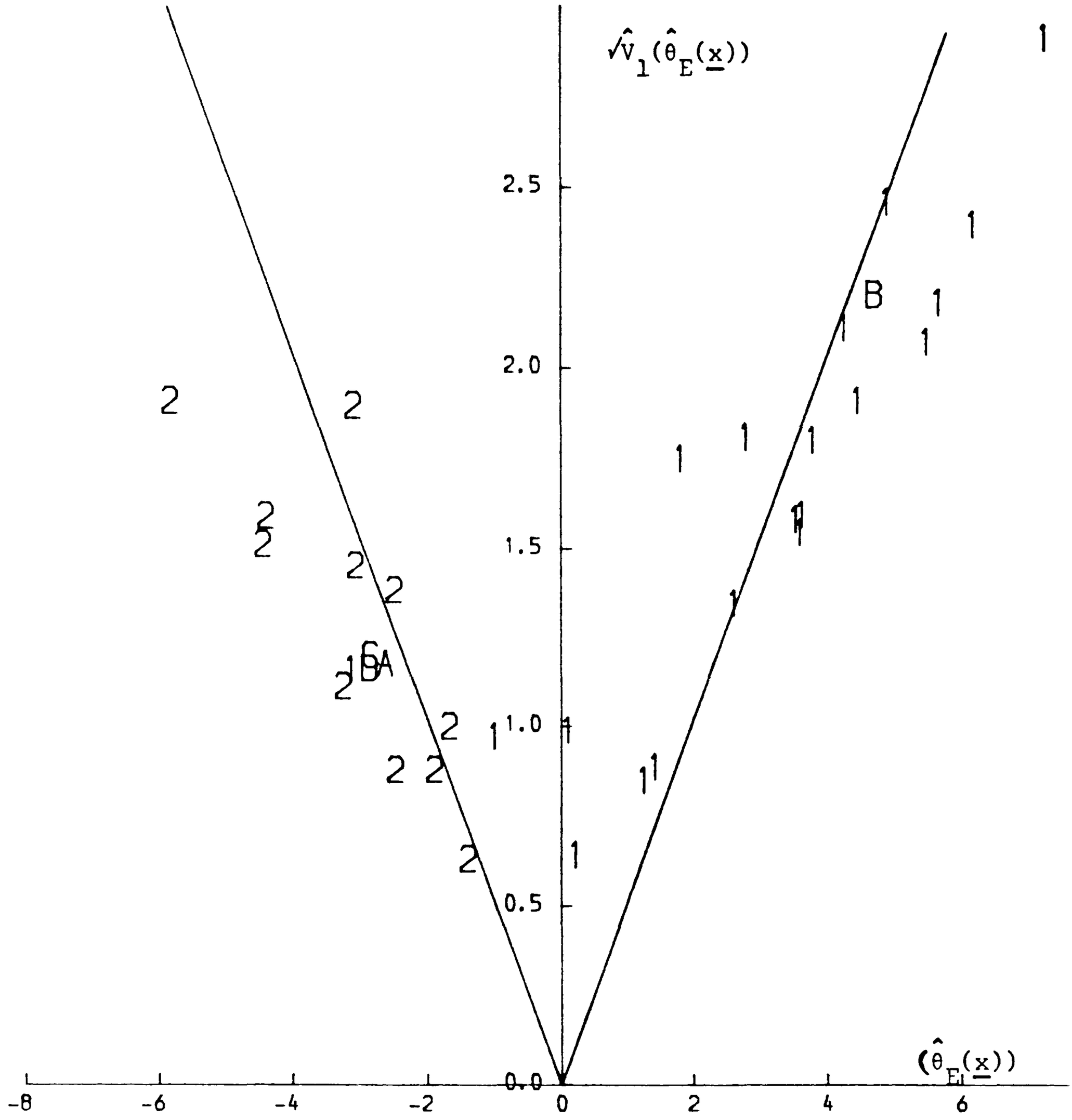


Figure (4(iv)b): CONN'S DATA

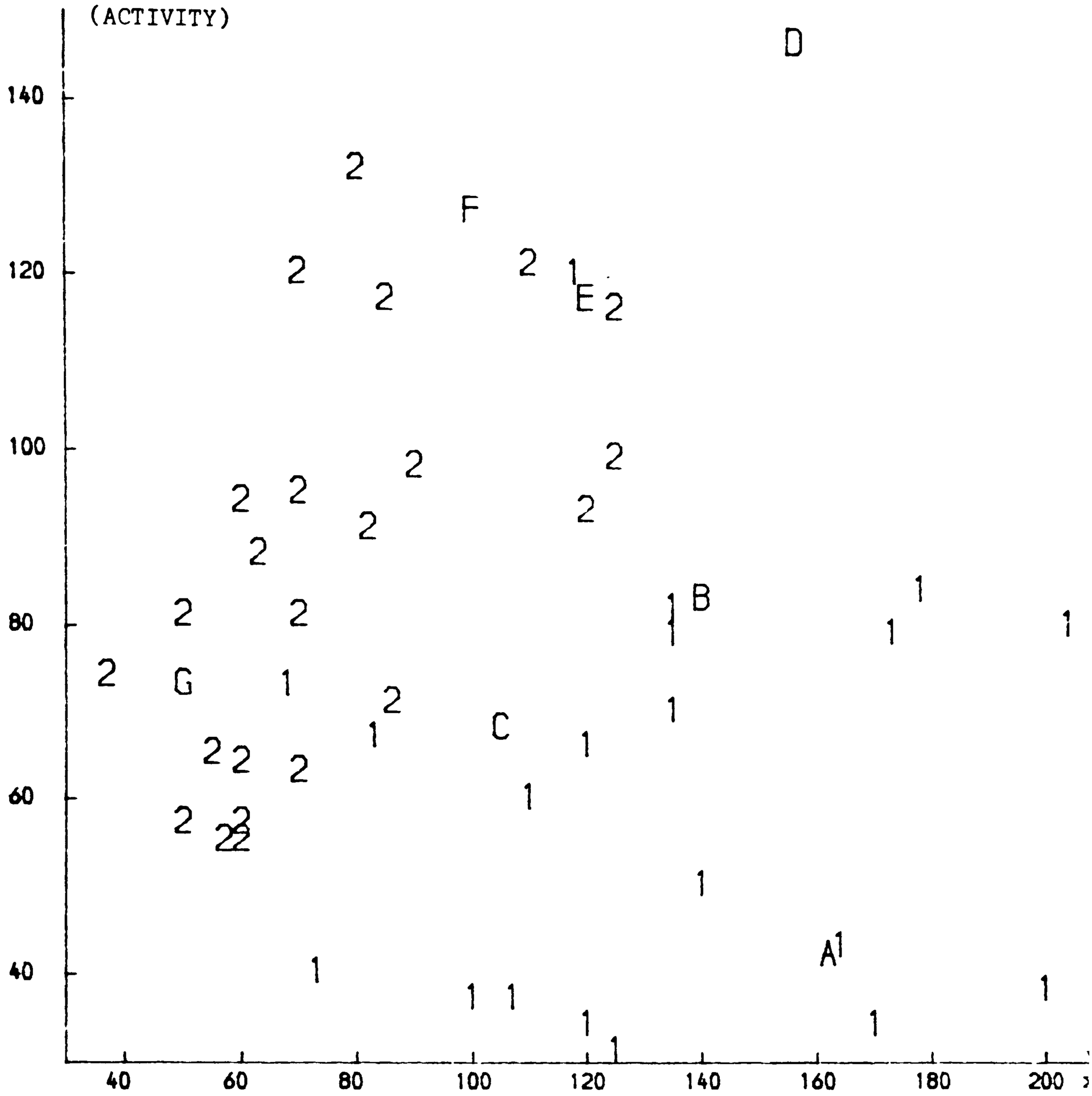


Figure (4(v)): HAEMOPHILIA DATA, SCATTER PLOT

1 ≡ Carrier

2 ≡ Control

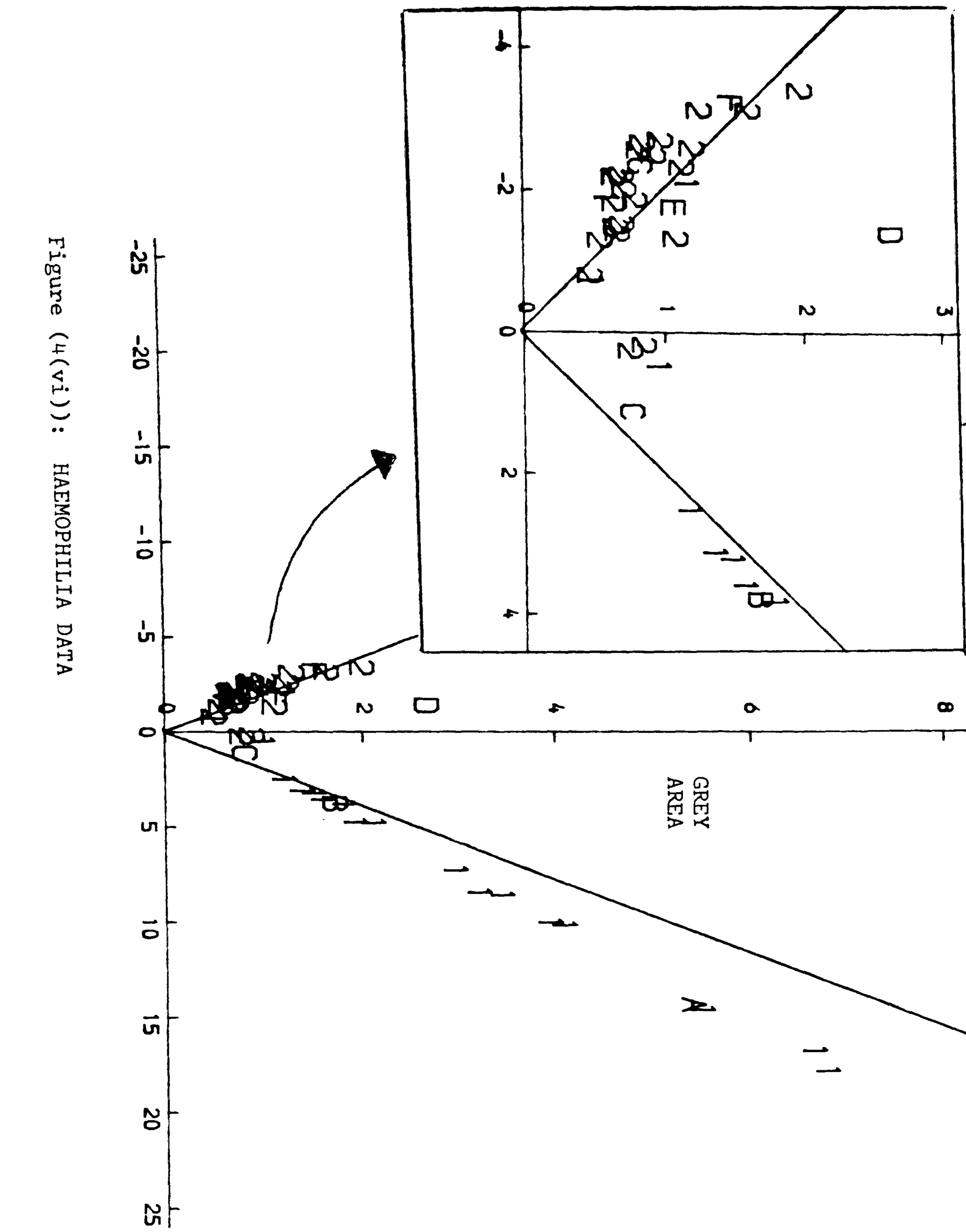


Figure (4(vi)): HAEMOPHILIA DATA

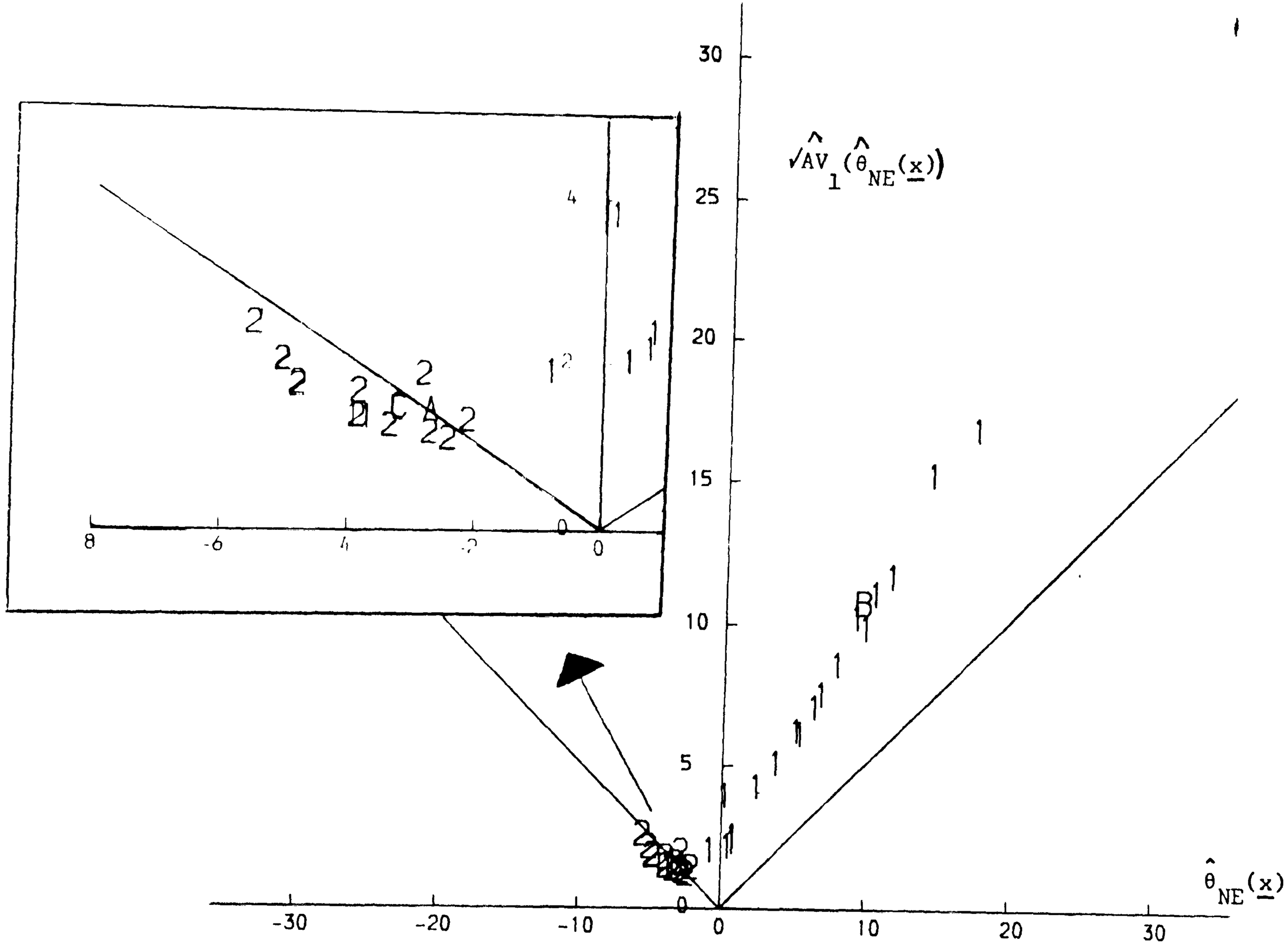


Figure (4(vii)): CONN'S DATA

CHAPTER 5

Alternative approaches for small sample sizes

5.1 Introduction

Figure 4(vii), of chapter 4, showed that using interval estimates of the type $(\hat{\theta}_{NE} \pm 1.96 \sqrt{\hat{AV}_1(\hat{\theta}_{NE})})$ did little to solve the discrimination problem for the CONN'S data, when assuming unequal covariance matrices. Assuming that more definite conclusions can be made for some patients, we consider possible reasons for this disappointing result.

(i) The derivation of $AV(\hat{\theta}_{NE}(\underline{x}))$ assumed a linear approximation of the Taylor series for $\hat{\theta}_{NE}(\underline{x})$ about $\hat{\underline{\beta}} = \underline{\beta}$ (see (2.3.3)).

(ii) The 'plug-in' estimate, $\hat{AV}_1(\hat{\theta}_{NE})$, see section (2.3), may be badly biased.

(iii) $\hat{\theta}_{NE}(\underline{x})$ itself is not approximately normally distributed.

We would expect that small sample sizes (relative to dimensionality) would make (i), (ii) and (iii) serious problems, and this is strongly suggested when we compare the discrimination results of the CONN's data and, say, the HAEMOPHILIA data (again assuming unequal covariances).

The sample sizes in the CONN'S DATA are even smaller than the smallest sample sizes considered in the simulation study of Chapter 3. In the simulation study we identifed significant non-normality of the distribution of $\hat{e}(\underline{x})$ in the small sample situation and hence we might expect this to be an important problem with the CONN'S DATA. Also, with respect to (i) above there was evidence in Chapter 3 that the approximation used in $AV(\hat{\theta}_{NE}(\underline{x}))$ was poorer in the small sample case.

In this chapter we consider other methods with the hope that they may be sensible alternatives (possibly approximate methods) to the techniques used in Chapter 4. We will look at:

(1) Replacing $AV_1(\hat{\theta}_{NE})$ by an unbiased estimator.

$AV_2(\hat{\theta}_{NE})$

(2) Bootstrap techniques

(3) The use of profile likelihood methods.

To illustrate the potential of these three methods, we will apply them to the CONN's data.

5.2 Unbiased Estimation

(A) Formula for unbiased estimator of $AV(\hat{\theta}_{NE})$

We consider first the problem of getting a better estimator for $AV(\hat{\theta}_{NE})$. An unbiased estimator of $AV(\hat{\theta}_{NE})$ could be a better estimator, and its derivation is as follows.

For each population Π_i.

$$f(\underline{x}|\Pi_i) \sim N_p(\underline{\mu}_i, \Omega_i) \quad (i = 1,2)$$

Let $\hat{\underline{\mu}}_i = \bar{\underline{x}}_i$, $\hat{\Omega}_i = \frac{1}{(n_i-p-2)} S_i$

where $S_i = \sum_{r=1}^{n_i} (\underline{x}_r - \bar{\underline{x}})(\underline{x}_r - \bar{\underline{x}})^T$

then $\frac{(n_i-p)}{p} \quad \frac{n_i}{(n_i-p-2)} (\underline{x}-\bar{\underline{x}}_i)^T \hat{\Omega}_i^{-1} (\underline{x}-\bar{\underline{x}}_i) \sim F(p, n_i-p, \lambda_i)$;

a non-central F distribution with non-centrality parameter

$$\lambda_i = n_i (\underline{x} - \underline{\mu}_i)^T \Omega_i^{-1} (\underline{x} - \underline{\mu}_i)$$

Let $\alpha_i^2(\underline{x}) = (\underline{x} - \underline{\mu}_i)^T \Omega_i^{-1} (\underline{x} - \underline{\mu}_i)$

and $\hat{\alpha}_i^2(\underline{x}) = (\underline{x} - \hat{\underline{\mu}}_i)^T \hat{\Omega}_i^{-1} (\underline{x} - \hat{\underline{\mu}}_i)$, which we will write as α_i^2 and $\hat{\alpha}_i^2$ to simplify the notation, $\quad E(\hat{\alpha}_i^2) = \frac{p}{n_i} + \alpha_i^2$.

$$E\left[(\hat{\alpha}_i^2)^2\right] = (\alpha_i^2)^2 \left[\frac{n_i-p-2}{n_i-p-4}\right] + \alpha_i^2 \left[\frac{2p}{n_i} + \frac{4(n_i-2)}{n_i(n_i-p-4)}\right]$$

$$+ \frac{2p(n_i-2)}{\quad} + \frac{p^2}{\quad}$$

Further, let $A_i = \frac{n_i-p-4}{n_i-p-2} (\hat{\alpha}_i^2)^2$

$$B_i = \hat{\alpha}_i^2 - \frac{p}{n_i}$$

We want an unbiased estimator for;

$$V_i = \frac{(\alpha_i^2)^2}{2(n_i-p-4)} + \left[\frac{1}{n_i} - \frac{(n_i - 2)}{(n_i-p-1)(n_i-p-4)}\right] \alpha_i^2$$

$$+ \frac{p(n_i - 2)}{2(n_i-p-1)(n_i-p-4)}$$

and where $AV\{\hat{\theta}_{NE}\} = V_1 + V_2$ (5.1.1)

Clearly, $E(A_i) = (\alpha_i^2)^2 + \alpha_i^2 \left[\frac{n_i-p-4}{n_i-p-2}\right] \left[\frac{2p}{n_i} + \frac{4(n_i-2)}{n_i(n_i-p-4)}\right]$

$$+ \frac{2p(n_i-2)}{n_i^2(n_i-p-2)} + \frac{p^2(n_i-p-4)}{n_i^2(n_i-p-2)}$$

and $E(B_i) = \alpha_i^2$

Finally, the unbiased estimator for V_i is $V_i' = \frac{A_i}{2(n_i-p-4)}$

$$+ B_i \left[\frac{1}{n_i} - \frac{(n_i-2)}{(n_i-p-1)(n_i-p-4)} - \frac{1}{2(n_i-p-2)}\left[\frac{2p}{n_i} + \frac{4(n_i-2)}{n_i(n_i-p-4)}\right]\right]$$

$$+ \frac{p(n_i-2)}{2(n_i-p-1)(n_i-p-4)} - \frac{p(n_i-2)}{n_i^2(n_i-p-2)(n_i-p-4)} - \frac{p^2}{2n_i^2(n_i-p-2)}$$

and $AV_2\{\hat{\theta}_{NE}\} = V_1' + V_2'$ .....(5.1.2)

(B) Application

The $(\hat{\theta}, \sqrt{var(\hat{\theta})})$ plot for the CONN's data, with $AV_2(\hat{\theta}_{NE})$ is given in figure [5(1)]. The unbiased estimator gives smaller estimated variances resulting in group 2 members all being out of the grey area, possibly an improvement over the use of $AV_1(\hat{\theta}_{NE})$. Unfortunately, we still have the initial problem with all of the group 1 patients still inside the GREY AREA.

We have indeed made some improvements in estimating the variance of $\hat{\theta}_{NE}$, using the unbiased estimator. However, the improvement was not enough to eliminate our current difficulty with

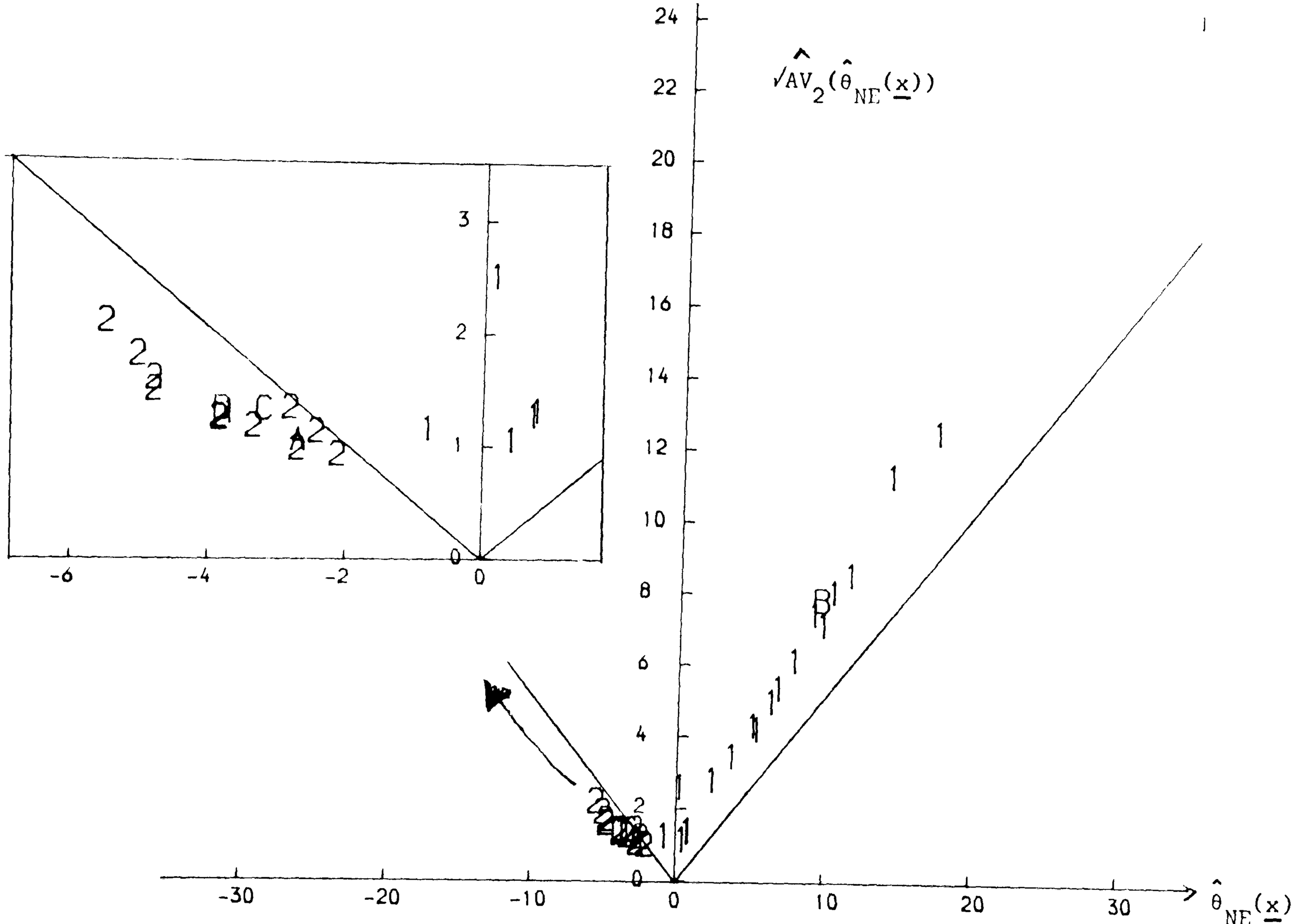


Figure (5(i)): CONN'S DATA

the CONN's data. Referring back to section 5.1, we suspect that using an approximate variance $AV(\hat{e}_{NE})$, could be the root of our problem.

The temptation here is to look for a better approximation to the variance of $\hat{e}_{NE}$, and hence a better estimator of $V(\hat{e}_{NE})$. However, the prospect of lengthy and tedious algebra, against getting a variance formula whose 'performance' is not guaranteed, strongly suggests looking at other ways of obtaining approximate interval estimates for e_{NE}. We will next look at the 'Bootstrap' technique.

5.3 BOOTSTRAP

(A) Description of the bootstrap

Given a data set, say $(\underline{y})$, we wish to study the distribution of $\hat{\theta}(\underline{y})$, which is the estimator of θ, the parameter of interest. Two problems may arise making theoretical results impossible to get.

(i) the distribution of y, F(y), is unknown

(ii) even if F(y) is known, the theory involved in obtaining the distribution of $\hat{\theta}(y)$ is difficult.

'Bootstrapping' [Efron (1982)] is an approximate method sometimes used to overcome these problems. It relies on raw computing power as a substitute for theoretical analysis.

An example of a bootstrap algorithm is as follows.

(i) Firstly we have the data set $y_1, y_2, \ldots, y_n$

(ii) Assign equal probabilities to y_i for all $i = 1, \ldots, n$ (i.e. $\hat{F}(y)$ is Multinomial).

(iii) Construct bootstrap sample $y_1^*, y_2^*, \ldots, y_n^*$.

Generate U(0,1) (i.e. random Uniform (0,1) variate)

If $0.0 \leq U < \frac{1}{n}$, $y(\ )^* = y_1$

$$\text{If } \frac{1}{n} \leq U < \frac{2}{n}, \quad Y_{(\ )}^{*} = Y_2$$

$$\vdots$$

$$\text{If } \frac{n-1}{n} \leq U < 1, \quad Y_{(\ )}^{*} = Y_n$$

Generating U n-times will give us our bootstrap-sample of size n. Given the boot-strap sample we can calculate $\hat{\theta}^{*}(\underline{Y}^{*})$.

(iv) Repeat (iii) B times and this gives $\hat{\theta}_1^{*}, \hat{\theta}_2^{*}, \ldots, \hat{\theta}_B^{*}$.

Let $\hat{\theta}^{*\bullet} = \dfrac{\sum_{i=1}^{B} \hat{\theta}_i^{*}}{B}$ and $\hat{SD}^{*} = \left[\dfrac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^{*}-\hat{\theta}^{*\bullet})^2\right]^{1/2}$

and as B tends to infinity, $\hat{\theta}^{*\bullet}$ tends to $\hat{\theta}(\underline{Y})$, and $\hat{SD}^{*}$ tend to the standard deviation where $F(y) = \hat{F}(y)$.

We stress the point that the bootstrap technique can only be regarded as an approximate method for our particular problem (CONN's data), since the data set is small ($n_1 = 20$, $n_2 = 11$, $p = 4$). We have to generate two bootstrap samples of size 20 and 11 respectively. As well as calculating $\hat{SD}^{*}$, we will consider other methods for constructing approximate interval estimates.

We consider various 'non-parametric' interval estimation methods suggested by Efron [1981, 1982 (Chapter 10)], Efron and Gong (1983, pg 40-41). Efron comments that "confidence intervals are often highly asymmetric about the best point estimate $\hat{\theta}$". To account for this asymmetry he suggests using the percentile method, which briefly is

Let $CDF(t) = \text{Prob}\{\hat{\theta}^{*} < t\}$

and estimate $CDF(t)$ by $\hat{CDF}(t) = \dfrac{\#\{\hat{\theta}_b^{*} \leq t\}}{B}$ [#≡ "number of times]

define $\hat{\theta}_{LOW}(\alpha) = \hat{CDF}^{-1}(\alpha)$, $\hat{\theta}_{UP}(\alpha) = \hat{CDF}^{-1}(1-\alpha)$, $0.0 < \alpha < 0.5$.

The percentile method then takes,

$$[\hat{\theta}_{LOW(\alpha)},\ \hat{\theta}_{UP(\alpha)}]$$

as an approximate $1-2\alpha$ central confidence interval.

Efron has reservations about the use of the percentile method. He considers other interval estimates, such as the "bias-corrected percentile method". Even so, in Efron and Gong (1983), Efron states that the appropriate theory justifying the use of these methods is still far from clear.

(B) Bootstrapping with the CONN's data

In Section (5.3. A) we did not make any specific assumption about F(y) and assigned equal probabilities to each data point. This could be considered a non-parametric bootstrapping method.

However for our problem, the formula for the log-odds θ is meaningful only if our data is multivariate normal, viz:- $\underline{X}_{\Pi_i} \sim Np(\underline{\mu}_i, \Omega_i)$, i=1,2. Efron calls this approach of assuming a parametric form for F(.) the "parametric bootstrap". We are immediately faced with the difficult question of how to estimate the parametric model to be used in the bootstrap. Efron (1982) gives no advice on this matter. To some extent we have considered the influence that different estimators for $(\underline{\mu}_i, \Omega_i)$ can have on the conclusions obtained. We considered alternative estimators for Ω_i, i.e.,

$$\frac{S_i}{(n_i-1)} \text{ and } \frac{S_i}{(n_i-p-2)},\ i=1,2. \quad \left[\begin{array}{l}\text{where } S_i \text{ is the corrected}\\ \text{sum of squares and cross}\\ \text{product matrix}\end{array}\right]$$

we note that both of these will result in $\hat{\theta}^*$ being a biased estimator for θ, though the second choice has been made to reduce this bias.

We note that in this problem the bootstrap distribution of $\hat{\theta}$ is obtained from estimating the distributions for Π_1, Π_2 separately and

sampling samples of size n_1 and n_2 for each population. See Figure (5(ii)). One-thousand bootstrap replications $[\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_{1000}]$ were obtained.

We studied two types of approximate 95% confidence interval for θ [note: θ referred to here is θ_{NE}].

(i) $\{\hat{\theta}^*_{(25)}, \hat{\theta}^*_{(975)}\}$, the percentile method. The $\hat{\theta}^*_{(.)}$ are order statistics of the bootstrap replications.

(ii) $\hat{\theta}^{*\bullet} \pm 1.96\ \hat{SD}^*$, with $\hat{\theta}^{*\bullet}$ and $\hat{SD}^*$ defined in section (5.3.1).

We note that the latter interval estimate is meaningful if $\hat{\theta}_i^*$ (i=1...,1000) is approximately normally distributed.

(C) Discussion of results for CONN's data

Firstly, in Figure (5(iii)) we have a histogram of $\hat{\theta}_i^*$ (i=1,...,1000) for points A and B (from test-set) using $\frac{1}{n_i-1} S_i$ as an estimator of Ω_i (i=1,2). In Table [5(i)] we list the interval estimates for θ together with $\hat{\theta}(\underline{x})$ itself.

$\hat{\Omega}_i$	interval estimate	A: $\hat{\theta}_{NE}(\underline{x}_A) = -2.71$	B: $\hat{\theta}_{NE}(\underline{x}_B) = 9.62$
$\frac{s_i}{n_i-1}$	$\hat{\theta}^{*\bullet} \pm 1.96\hat{SD}^*$	-7.70, 2.90	-17.55, 62.0
	$\hat{\theta}^*_{(25)}, \hat{\theta}^*_{(975)}$	-6.80, 3.30	3.0, 73.0
$\frac{s_i}{n_i-p-2}$	$\hat{\theta}^{*\bullet} \pm 1.96\hat{SD}^*$	-5.38, 1.24	-13.15, 34.25
	$\hat{\theta}^*_{(25)}, \hat{\theta}^*_{(975)}$	-4.60, 1.30	0.50, 37.50

Table [5(I)]: "Parametric bootstrapping" for point A and B of CONN's data.

The histogram for point A is more symmetric. Point B's histogram is considerably more skewed and has larger outliers in its upper tail. More of the $\hat{\theta}_i^*$ for point A have negative values.

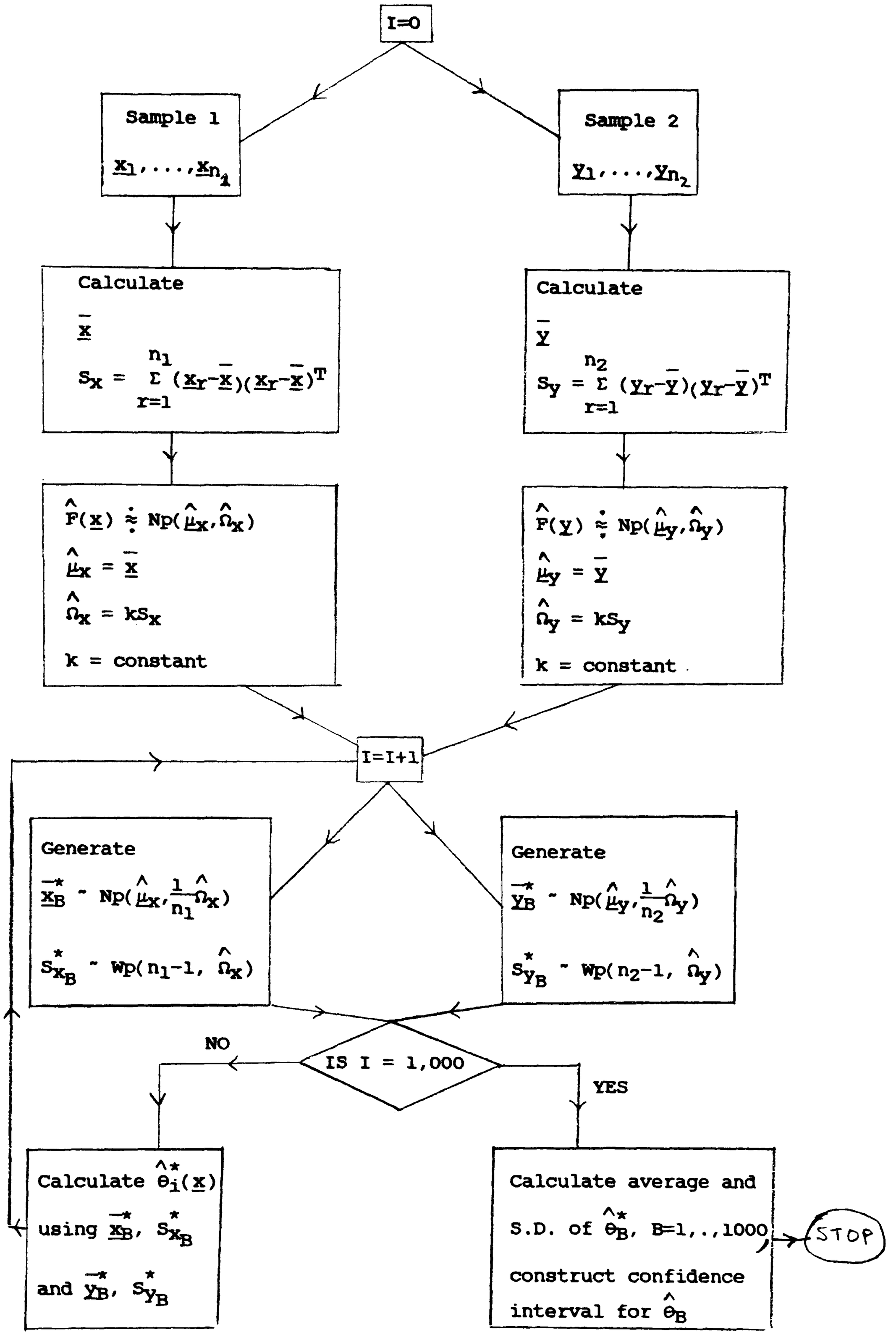


FIGURE (5(ii)) {Note: $\hat{\theta}^*(\underline{X}) = \hat{\theta}_{NE}(\underline{X})$}

```
EACH * REPRESENTS  10 OBSERVATIONS

MIDDLE OF    NUMBER OF
INTERVAL     OBSERVATIONS
 -15.00        2     *
 -13.00        0
 -11.00        0
  -9.00        7     *
  -7.00       31     ****
  -5.00      160     ****************
  -3.00      453     *********************************************
  -1.00      237     ************************
   1.00       65     *******
   3.00       25     ***
   5.00        5     *
   7.00        7     *
   9.00        3     *
  11.00        3     *
  13.00        0
  15.00        0
  17.00        0
  19.00        0
  21.00        1     *
  23.00        0
  25.00        0
  27.00        1     *


EACH * REPRESENTS  10 OBSERVATIONS
1 OBSERVATIONS ARE BELOW THE FIRST CLASS

MIDDLE OF    NUMBER OF
INTERVAL     OBSERVATIONS
    0.0        49     *****
   10.0       394     ****************************************
   20.0       257     **************************
   30.0       135     **************
   40.0        77     ********
   50.0        35     ****
   60.0        17     **
   70.0        14     **
   80.0         7     *
   90.0         5     *
  100.0         1     *
  110.0         1     *
  120.0         1     *
  130.0         0
  140.0         1     *
  150.0         0
  160.0         2     *
  170.0         0
  180.0         2     *
  190.0         0
  200.0         0
  210.0         0
  220.0         1     *
```

Figure (5(iii)): CONN'S DATA. Top most stem-and-leaf plot for point (A), and for point (B) the plot below. For both points $\hat{\Omega}_i = S_i/(n_i-1)$.

From table [5(I)], the percentile interval for point A captures zero and the magnitudes of $\hat{e}_i^*$ are small (note $\hat{e}_{NE}(\underline{x}_A) = -2.71$). We would be very uncertain of classifying point A to population two. For point B, the percentile interval does not capture zero and is skewed towards large positive values ($\hat{e}_{NE}(\underline{x}_B) = 9.62$). The percentile interval does reasonably well in capturing the shape of the distribution of $\hat{e}_i^*$. We would therefore be fairly confident in saying that point B is a group 1 member. The results of the percentile method contradict the conclusions we would make (for points A and B) from figure [5(I)], i.e. the $(\hat{e}, \sqrt{var(\hat{e})})$ plot.

All the intervals give essentially the same results for point A. This is largely due to the symmetry of the distribution of $\hat{e}_i^*$. For point B, the interval $(\hat{e}^{*\bullet} \pm 1.96\hat{SD}^*)$ captures zero and contains large negative values. This is a direct result of the asymmetry of the distribution of $\hat{e}_i^*$, clearly $\hat{e}_i^*$ is not normally distributed.

We also note that the choice of $\hat{\Omega}_i$ did not change the general conclusions for both types of interval estimates. We do however notice that using $S_i/[n_i-p-2]$ gives narrower interval estimates.

From the various comments of Efron (1982), much more could be exploited from the bootstrap method. However, since the whole subject of Bootstrapping is still in an "exploratory" state, we leave this technique for the moment, optimistic of the potential it holds for small-sample discrimination problems.

We proceed next to another alternative technique based on the profile likelihood.

5.4 Profile-Likelihood ($\Omega_1 \neq \Omega_2$)

(A) Introduction

Using the notation from section 2.1, the density function of $\underline{x}$ is,

$$f(\underline{x}|\Pi_i) \sim Np(\underline{\mu}_i, \Omega_i) \quad (i=1,2)$$

Let the elements of $\underline{\eta}$ be the distinct elements of μ_1, Ω_1, μ_2, Ω_2. Define the likelihood function as Lik(θ, $\underline{\eta}$).

We use this notation for the likelihood for convenience. Note that for given $\underline{\eta}$, θ is a redundant parameter.

Further, let $\ell(\theta, \underline{\eta}) = \log [\text{Lik}(\theta, \underline{\eta})]$.

One method of eliminating $\underline{\eta}$ from the likelihood function is to maximise over it for each θ.

$$\text{let } \ell[\theta, \hat{\underline{\eta}}(\theta)] = \max_{\underline{\eta}} [\ell(\theta, \underline{\eta})]$$

Using the ideas of Kalbfleisch (1979) we can describe $\ell(\theta, \hat{\underline{\eta}}(\theta))$ as the 'profile' or 'silhouette' of the log likelihood function when viewed over the θ-axis. Kalbfleisch (1979) describes the 'profile' of the 'maximum relative likelihood'. In Kalbfleisch and Sprott (1970), the 'maximum relative likelihood' is one of four likelihood methods for eliminating large numbers of nuisance parameters. These authors suggest using the following interval estimate for θ,

$$IE_\theta = \{\theta: \ell(\theta, \hat{\underline{\eta}}(\theta)) - \ell(\hat{\theta}, \hat{\underline{\eta}}(\theta)) > h\} \qquad (5.4.1)$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ. From large sample results, an approximate 95% interval estimate for θ may be obtained by letting h equal $-0.5\ \chi^2(1;\ c)$, where $c = 0.95$. We note that IE_θ must be regarded cautiously for small samples.

(B) Derivation of the profile likelihood

For each $\theta = C$ say we must find the maximising $\underline{\eta}$. We have a constrained maximisation problem, viz:-

$$\max_{\underline{\eta}} \ell(\theta, \underline{\eta}) \text{ subject to } \theta = C$$

Introduce the Lagrangian function L and the Lagrangian multiplier λ.

For typographic convenience we shall write $\ell(\theta, \underline{\eta})$ as $\ell(\underline{\eta})$.

We therefore want the stationary points of

$$L(\underline{\eta}, \lambda) = \ell(\underline{\eta}) + \lambda(\theta - C)$$

The solution is a saddle-point of L and we:

(i) Maximise L w.r.t. $\underline{\eta}$. Denote this maximum by $\hat{\underline{\eta}}(\lambda)$.

(ii) Substitute $\hat{\underline{\eta}}(\lambda)$ into L.

(iii) Minimise $L[\hat{\underline{\eta}}(\lambda), \lambda]$ w.r.t. λ.

For purposes of notation $\hat{\underline{\eta}}(\lambda)$ and $\hat{\underline{\eta}}(\theta)$ are used, their interpretation is clear within the context.

For a discussion of steps (i), (ii) and (iii) above, see Whittle (1971) chapter 3.

It can be shown that,

$$\ell(\underline{\eta}) = \frac{-n_1}{2}\log|\Omega_1| - \frac{1}{2}\sum_j(\underline{x}_{1j}-\underline{\mu}_1)^T\Omega_1^{-1}(\underline{x}_{1j} - \underline{\mu}_1)$$

$$- \frac{n_2}{2}\log|\Omega_2| - \frac{1}{2}\sum_j(\underline{x}_{2j}-\underline{\mu}_2)^T\Omega_2^{-1}(\underline{x}_{2j}-\underline{\mu}_2) + \text{constant},$$

where $\underline{x}_{kj}$ is observation j from population k, (k=1,2) and,

$$\lambda(\theta(\underline{x}) - C) = -\frac{\lambda}{2}\log|\Omega_1| - \frac{\lambda}{2}(\underline{x}-\underline{\mu}_1)^T\Omega_1^{-1}(\underline{x}-\underline{\mu}_1)$$

$$+ \frac{\lambda}{2}\log|\Omega_2| + \frac{\lambda}{2}(\underline{x}-\underline{\mu}_2)^T\Omega_2^{-1}(\underline{x}-\underline{\mu}_2) - \lambda C.$$

Note that $\ell(\underline{\eta}) + \lambda(\theta-C)$ is 'like' a log-likelihood with $\underline{x}$ added to sample 1 with weight λ and with $\underline{x}$ added to sample 2 with weight $-\lambda$. From the usual likelihood estimates for normally distributed data, we get the following unique parameter estimates.

$$\left.\begin{aligned}
\hat{\underline{\mu}}_1 &= \frac{n_1\bar{\underline{x}}_1 + \lambda\underline{x}}{(n_1 + \lambda)}, \quad \hat{\underline{\mu}}_2 = \frac{n_2\bar{\underline{x}}_2 - \lambda\underline{x}}{(n_2 - \lambda)} \\
\hat{\Omega}_1 &= \frac{1}{(n_1+\lambda)}\left[S_1 + \frac{\lambda n_1}{(\lambda+n_1)}(\underline{x} - \bar{\underline{x}}_1)(\underline{x} - \bar{\underline{x}}_1)^T\right] \\
\hat{\Omega}_2 &= \frac{1}{(n_2-\lambda)}\left[S_2 - \frac{\lambda n_2}{(n_2-\lambda)}(\underline{x} - \bar{\underline{x}}_2)(\underline{x} - \bar{\underline{x}}_2)^T\right]
\end{aligned}\right\} \quad (5.4.2)$$

$$L(\hat{\underline{\eta}}(\lambda), \lambda) = \frac{-(n_1+n_2)}{2}p\log(2\Pi)$$

$$+\frac{(n_1 + \lambda)}{2}p\log(n_1+\lambda) \qquad +\frac{(n_2 - \lambda)}{2}p\log(n_2-\lambda)$$

$$\frac{-(n_1+\lambda)}{2}\log|S_1| \qquad \frac{-(n_2-\lambda)}{2}\log|S_2|$$

$$\frac{-(n_1+\lambda)}{2}\log\left[1+\frac{\lambda n_1}{(\lambda+n_1)}a_1\right] \quad \frac{-(n_2-\lambda)}{2}\log\left[1-\frac{\lambda n_2}{(n_2-\lambda)}a_2\right]$$

$$-\frac{1}{2}(n_1+n_2)p - \lambda C \qquad (5.4.3)$$

where $a_i = (\underline{x} - \bar{\underline{x}}_i)^T S_i^{-1}(\underline{x} - \bar{\underline{x}}_i)$

Minimization of $L(\hat{\underline{\eta}}(\lambda),\lambda)$ must be carried out numerically. We used the Newton-Raphson method of numerically obtaining the solution of $\frac{\delta L}{\delta\lambda} = 0$. We have,

$$\frac{\delta L}{\delta\lambda} = \frac{(p+1)}{2}\left[\log\left[\frac{n_1+\lambda}{n_2-\lambda}\right]\right] - \frac{1}{2}\log|S_1| + \frac{1}{2}\log|S_2| - C$$

$$- \frac{1}{2}\log\left[\frac{n_1+\lambda+\lambda n_1 a_1}{n_2-\lambda-\lambda n_2 a_2}\right] - \frac{1}{2}\left[\frac{n_1^2 a_1}{\lambda+n_1+\lambda n_1 a_1}\right]$$

$$+ \frac{1}{2}\left[\frac{n_2^2 a_2}{n_2-\lambda-\lambda n_2 a_2}\right] \qquad (5.4.4)$$

From (5.4.3) and (5.4.4) we get constraints on λ.

(i) $n_1 + \lambda > 0$

(ii) $1 + \frac{\lambda n_1 a_1}{(n_1+\lambda)} > 0$

(iii) $n_2 - \lambda > 0$

(iv) $1 - \frac{\lambda n_2 a_2}{(n_2-\lambda)} > 0$

(i), (ii), (iii) and (iv) are equivalent to,

$$\frac{-n_1}{(n_1 a_1 + 1)} < \lambda < \frac{n_2}{(n_2 a_2 + 1)} \qquad (5.4.5)$$

We have a unique solution to the minimisation since we can show that $\frac{\delta^2 L}{\delta\lambda^2} > 0$.

$$\frac{\delta^2 L}{\delta\lambda^2} = \frac{(p+1)}{2}\left[\frac{1}{n_1+\lambda} + \frac{1}{n_2-\lambda}\right]$$

$$\frac{1}{2}\frac{(n_1 a_1+1)}{(\lambda+n_1+\lambda n_1 a_1)} \qquad \frac{1}{2}\frac{(n_2 a_2+1)}{(n_2-\lambda-\lambda n_2 a_2)}$$

$$+\frac{1}{2}\ \frac{n_1^2 a_1(n_1a_1+1)}{(\lambda+n_1+\lambda n_1a_1)^2}\qquad +\frac{1}{2}\ \frac{n_2^2 a_2(n_2a_2+1)}{(n_2-\lambda-\lambda n_2a_2)^2}$$

$$=\frac{1}{2}\left[\frac{p(\lambda+n_1+\lambda n_1a_1)^2+n_1^4a_1^2}{(n_1+\lambda)(\lambda+n_1+\lambda n_1a_1)^2}+\frac{p(n_2-\lambda-\lambda n_2a_2)^2+n_2^4a_2^2}{(n_2-\lambda)(n_2-\lambda-\lambda n_2a_2)^2}\right]$$

Clearly $\frac{\delta^2 L}{\delta\lambda^2} > 0$, i.e. the function is convex, for λ given in (5.4.5).

The question arises as to whether we have in fact solved the constrained optimisation problem we set out to solve. We can always check whether we have the correct solution by testing that the constraint is satisfied at the values of $\hat{\underline{\mu}}_1$, $\hat{\underline{\mu}}_2$, $\hat{\Omega}_1$, $\hat{\Omega}_2$ (i.e. (5.4.2)). In later examples, we have empirically found that the constraint is always satisfied at our solution. In Appendix (5.2) we offer a heuristic argument that the fact that $\hat{\underline{\mu}}_1$, $\hat{\underline{\mu}}_2$, $\hat{\Omega}_1$, $\hat{\Omega}_2$ in (5.4.2) are unique will imply that we have the correct solution to our constrained optimisation problem.

(C) Application to the CONN's data

For the CONN's data $\ell(\hat{\theta}, \hat{\underline{\eta}}(\hat{\theta})) = 41.037$ and $h = -1.92$. The suggested approximate interval estimate for θ, by (5.4.1) is,

$$IE_\theta = \{\theta: \ell(\theta, \hat{\underline{\eta}}(\theta)) > 39.12\}$$

Figures (5(iv)a) and (5(iv)b) are the plots of the profile of the log-likelihood for points A and B of the CONN's data test set. Each point on those plots is the numerical solution to the constrained maximisation problem for a given "$\theta=C$". In these examples we have checked that the constraint $\theta=C$ is always satisfied.

The graph for point A is more symmetric than for point B. This may suggest using either of IE_θ or $\{\hat{\theta} \pm 1.96\sqrt{var(\hat{\theta})}\}$ for point A, and perhaps use only IE_θ for point B.

For point B, IE_θ excludes zero. For point A the values of θ in the interval are relatively small, and IE_θ captures zero. Comparing

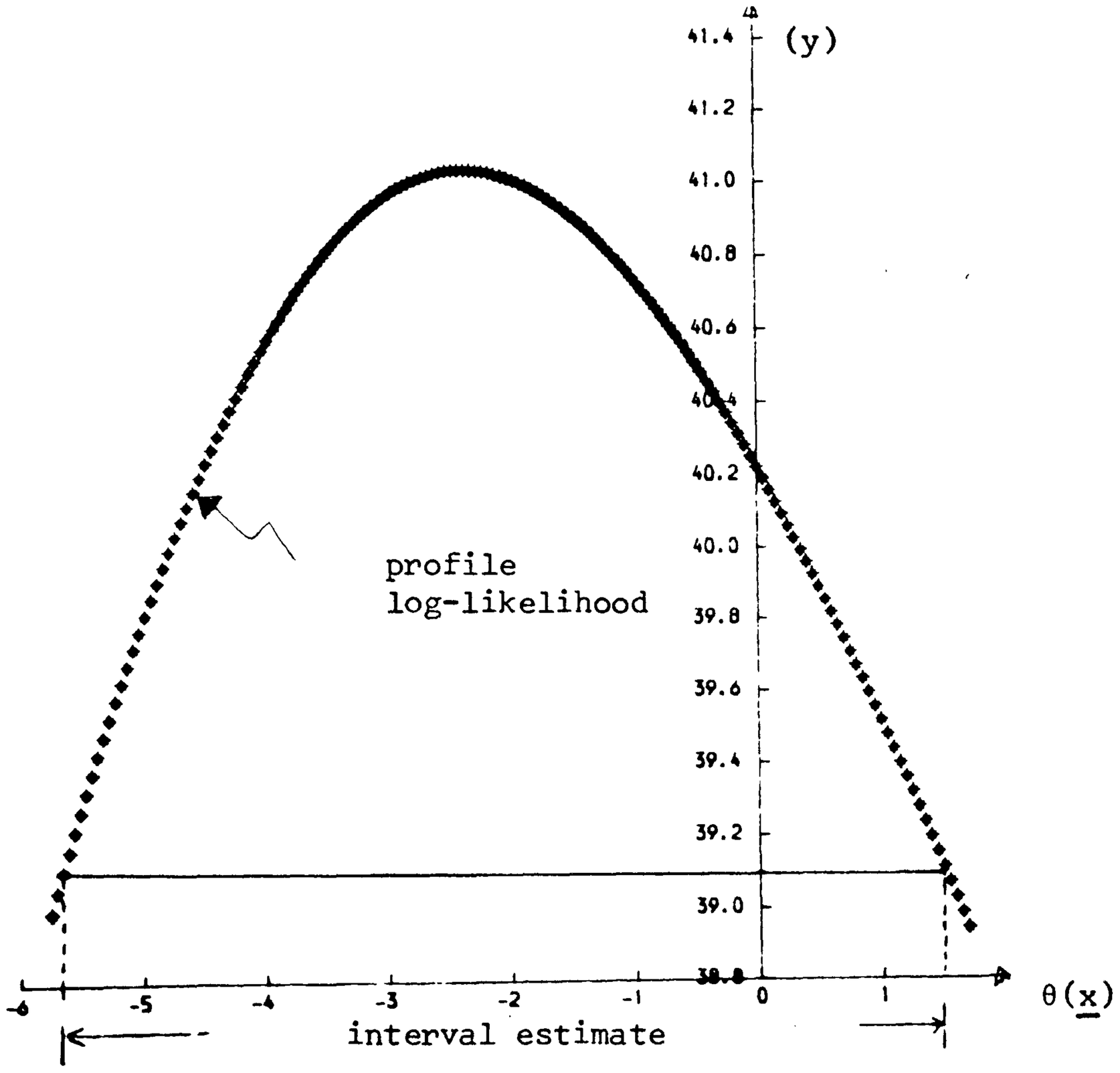


Figure (5(iv)a): CONN'S DATA, POINT (A)

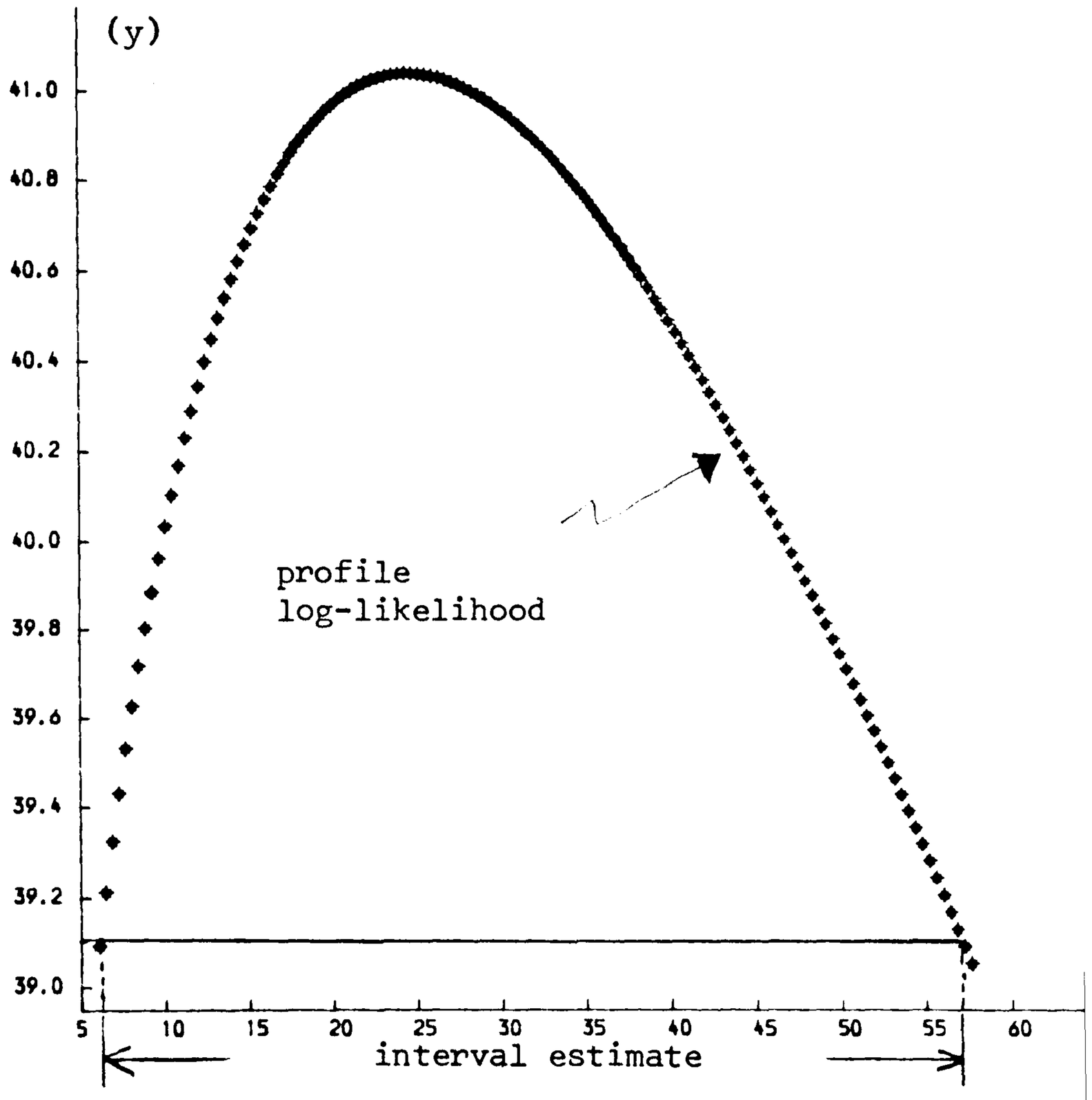


Figure (5(iv)b): CONN'S DATA, POINT (B)

IE_θ for point B with the $(\hat{\theta},\ SD(\hat{\theta}))$ plot (in figure 5(i)) the profile likelihood method does seem promising with respect to assigning group one members with some certainty.

(5.5.) Comparison of the three methods using CONN's data

We compare all three methods discussed so far by looking at Table [5(ii)] which gives the interval estimates for the various methods; for points A, B, C, D of the CONN's DATA test set.

We first compare the methods over all four points A, B, C and D. The comparison is made in terms of the allocation process. Intervals using $\hat{AV}_2(\hat{\theta}_{NE})$ gave the same result as those using $\hat{AV}_1(\hat{\theta}_{NE})$. For a detailed comparison of these two formulae refer to (2.3.1) and (5.1.2). $(\hat{\theta}^*,\ \pm\ 1.96\ \hat{SD}^*)$ gave the same result for both estimators of Ω_i. We note that $\hat{SD}^*$ is typically larger than $(\hat{AV}_i(\hat{\theta});\ i=1,2)$. Looking at figure [5(iii)] we can see that $\hat{\theta}^*$ is not normally distributed having long tailed distributions which may make $\hat{SD}^*$ large. These features are probably due to the small sample sizes in this problem. On the other hand, construction of $(\hat{AV}_i(\hat{\theta});\ i=1,2)$ is based on a large sample approximation.

The intervals from the percentile method (both $\hat{\Omega}_i$) and the profile likelihood method gave the same results. They can however differ from either or both of $(\hat{\theta}\ \pm\ 1.96\sqrt{\hat{AV}_i(\hat{\theta})};\ i=1,2)$ and $(\hat{\theta}^*,\ \pm\ 1.96\hat{SD}^*)$.

We now look at specific $\underline{x}$-points and compare all the methods.

CONN'S DATA (unclassified points)				
Approx. Var $\hat{\theta}_{NE} \pm 1.96\ \hat{SD}(\hat{\theta})$	A	B	C	D
	$\hat{\theta}_{NE}(x_A) = -2.71$	$\hat{\theta}_{NE}(x_B) = 9.62$	$\hat{\theta}_{NE}(x_C) = -3.22$	$\hat{\theta}_{NE}(x_D) = -3.83$
$\hat{AV}_1(\hat{\theta}_{NE})$	-5.42, -0.00	-11.13, 30.38	-5.99, -0.44	-6.40, -1.26
$\hat{AV}_2(\hat{\theta}_{NE})$	-4.72, -0.69	-5.43, 24.68	-5.77, -0.67	-6.38, -1.28
Parametric Bootstrap $\hat{\Omega}_i = S_i/(n_i-1)$				
$\hat{\theta}^{*\cdot} \pm 1.96\ \hat{SD}^*$	-7.70, 2.90	-17.55, 62.03	-7.95, 1.66	-8.00, -1.26
$[\hat{\theta}^*_{(25)}, \hat{\theta}^*_{(975)}]$†	-6.80, 3.30	3.0, 73.0	-7.70, 2.30	-8.50, -1.8
Parametric Bootstrap $\hat{\Omega}_i = S_i/(n_i-p-2)$				
$\hat{\theta}^{*\cdot} \pm 1.96\ \hat{SD}^*$	-5.38, 1.24	-13.15, 34.29	-5.74, 0.45	-5.88, -0.75
$[\hat{\theta}^*_{(25)}, \hat{\theta}^*_{(975)}]$	-4.60, 1.30	0.50, 37.50	-5.80, 0.40	-6.30, -1.30
PROFILE LIKELIHOOD	-5.65, 1.50	6.17, 57.03	-6.69, 0.42	-7.63, -2.38

TABLE (5(ii))

(†) $\hat{\theta}^*_{(.)}$ are the order statistics $\hat{\theta}^*_{(25)}$ $\hat{\theta}^*_{(975)}$ are the 0.025 and 0.975 quantiles.

$\underline{x}$-point	$\hat{\theta}(\underline{x})$	$\hat{AV}_2\{\hat{\theta}_{NE}\}$	interval estimate for θ using profile likelihood
$\underline{x}_5$	-3.716	1.643	(-7.40, -2.2)
$\underline{x}_8$	34.777	1005.523	(26.0, ..?..)
$\underline{x}_9$	7.862	71.239	(5.0, 46.5)
$\underline{x}_{18}$	2.376	16.814	(0.4, 21.40)

TABLE (5(iii)) CONN'S (GROUP 1)

Note (i) The plots of the profile likelihood are given in figures [5(v)a] to [5(v)d]

(ii) The question mark (..?..) for $\underline{x}_8$ indicates that an upper limit for the interval estimate was not found.

The percentile and profile methods give positive intervals for point B, but they include zero for the other methods.

Points A and C have wholly negative intervals with $(\hat{\theta}_{NE} \pm 1.96\ \hat{AV}_i(\hat{\theta}_{NE}), i=1,2)$ while for the other methods the intervals capture zero. We further note that the upper boundary of the intervals using $\hat{AV}_i(\hat{\theta}_{NE})$, for points A and C are close to zero. Hence despite this difference in overlap with zero all of the intervals are similar.

Only for point D do all methods clearly agree. This may be due to the distribution of $\hat{\theta}$ being approximately normal and the profile $\log_\lambda$-likelihood being quadratic at this value of $\underline{x}$.

All methods tend to give similar results for point A, but this is not true for point B. This difference is probably due to the distribution of $\hat{\theta}$ [figure 5(iii)] and the profile likelihood [figure 5(iv)a and 5(iv)b] for point A being more symmetric than it is for point B.

An interesting result here is that, the profile and percentile methods do allocate point B, with some certainty, to group 1. This suggests that either of these two methods would classify more of the group 1 points with some level of certainty.

We now use the profile likelihood method on some other group 1 members. The preference over the bootstrap method is that the profile likelihood method is computationally less time consuming, and all we require are:

$$a_i = (\underline{x} - \bar{\underline{x}}_i)^T S_i^{-1} (x - \bar{\underline{x}}_i) \text{ and } \log|S_i| \quad (i=1,2).$$

The following four points are chosen. The points $\underline{x}_5$ and $\underline{x}_{18}$ have small values for $\hat{\theta}(\underline{x})$, but with opposite signs. $\underline{x}_9$ has a relatively large $\hat{\theta}(\underline{x})$. $\underline{x}_8$ is an outlying value. Table [5(iii)] gives the values of $\hat{\theta}$, $\hat{AV}_2(\hat{\theta})$, and the interval estimate for θ, for each of these four points.

For $\underline{x}_8$, numerical difficulties were encountered in finding the upper limit of the interval estimate. We see, from figure (5(v)b), that the upper bound of the interval will be very large. We would expect the terms $a_i = (\underline{x} - \overline{\underline{x}_i})^T S_i^{-1} (x - \overline{\underline{x}_i})$ to be large also. From (5.4.5) large a_i-values would make λ reach its boundary values, and from (5.4.3) $L(\hat{\underline{\eta}}(\lambda), \lambda)$ will no longer be quadratic. In this situation the Newton-Raphson method will take us outside the range of feasible λ. We did not persevere with this numerical problem considering that the interval estimate for $\theta(\underline{x}_8)$ is clearly very wide and it is also wholly positive.

All the points are correctly classified except $\underline{x}_5$. Looking back at the $(\hat{\theta}, SD(\hat{\theta}))$ plot in figure (5(i)) the point $\underline{x}_5$ is clearly inside the cluster of group 2 members and would possibly be misclassified by any method. Correctly classifying $\underline{x}_8$, $\underline{x}_9$ and $\underline{x}_{18}$ suggests that we could also correctly classify most of the other group 1 members.

(5.6) Summary

We compared three approximate interval estimation methods for θ: using $(AV_i(\hat{\theta})$; i=1,2), the bootstrap and profile likelihood methods on the CONN's data.

The profile and 'percentile' bootstrap methods tend to classify the points in the same way. Both have wide interval estimates for θ for group 1 members, and the corresponding intervals for group 2 are narrower. Using $(\hat{\theta} \pm 1.96\sqrt{Var(\hat{\theta})})$ can give misleading results due to the non-normality of the distribution of $\hat{\theta}$. We see this, for example, point B in Table [5(ii)] when we used $(\hat{\theta}_{NE} \pm 1.96\sqrt{AV_i(\hat{\theta}_{NE})})$ (i=1,2) and $(\hat{\theta}^{*\bullet} \pm 1.96\hat{SD}^{*})$.

Group 2 $\underline{x}$-points are more tightly clustered [e.g. figure (5(i))] so all $\alpha_1^2(\underline{x})$ will be similar and not too large. We might therefore expect all 3 methods to perform similarly on group 2 members. Point D in Table [5(ii)] is a good example. Points A and C

(Table 5(II)) have similar intervals for all methods (see section 5.5).

To estimate e, we need to estimate the atypicality terms, $\alpha_i^2(\underline{x}) = (\underline{x}-\underline{\mu}_i)^T\Omega_i^{-1}(\underline{x}-\underline{\mu}_i)$. Clearly group 1 members have large atypicality with respect to population 2. Further the variance of $\hat{e}_{NE}(\underline{x})$ increases with $\alpha_i^2(\underline{x})$, and large $\alpha_i^2(\underline{x})$ will mean large interval estimates. We also note that group 2 members are not very atypical with respect to population 1. Therefore, we can expect group 1 members to have a wider range of $\hat{e}(\underline{x})$, and often wider interval estimates for e when compared to group 2 members.

All the intervals in Table [5(II)] give a large range for the odds. Consider for example point C, using the profile likelihood method. We have on one extreme, odds of 804:1 of being in group 2 ($\simeq e^{-6.69}$) and on the other extreme an odds of 1.5:1 of being in group 1 ($\simeq e^{0.42}$). These extreme odds are much further apart for point B in group 1. Clearly for sample sizes as small as these we should be very uncertain of allocating an $\underline{x}$-point based only on its point estimate $\hat{e}_{NE}(\underline{x})$. However, although the true value of the odds will be very uncertain, decisions can still be made, e.g. for point B using the profile likelihood method whose interval estimate for e is wholly positive.

With respect to allocating group 1 members with some certainty, the results suggest constructing interval estimates for e with the profile likelihood or percentile method. However, we emphasise that the conclusions drawn, from the application of these small sample techniques to a single problem, should be treated cautiously.

Throughout our study, so far, we have explicitly assumed normality of the distribution of $\underline{x}$. In the next chapter we briefly look at the consequences of non-normality on our approach to the two population discrimination problem.

For completeness the relevent formulae for the equal covariance case, when using the profile likelihood method is given in Appendix (5.1).

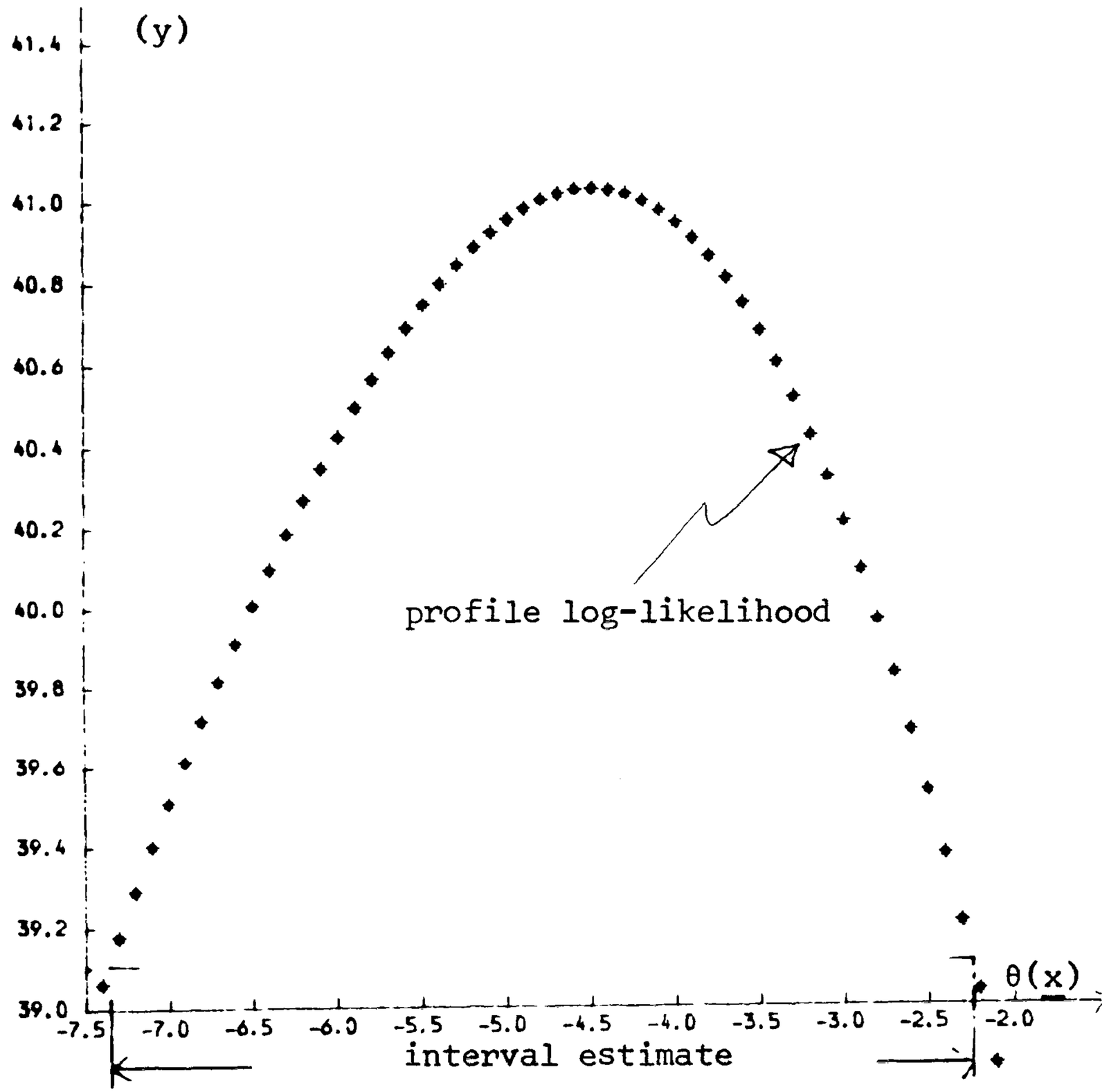


Figure (5(v)a): CONN'S DATA, POINT $\underline{x}_5$

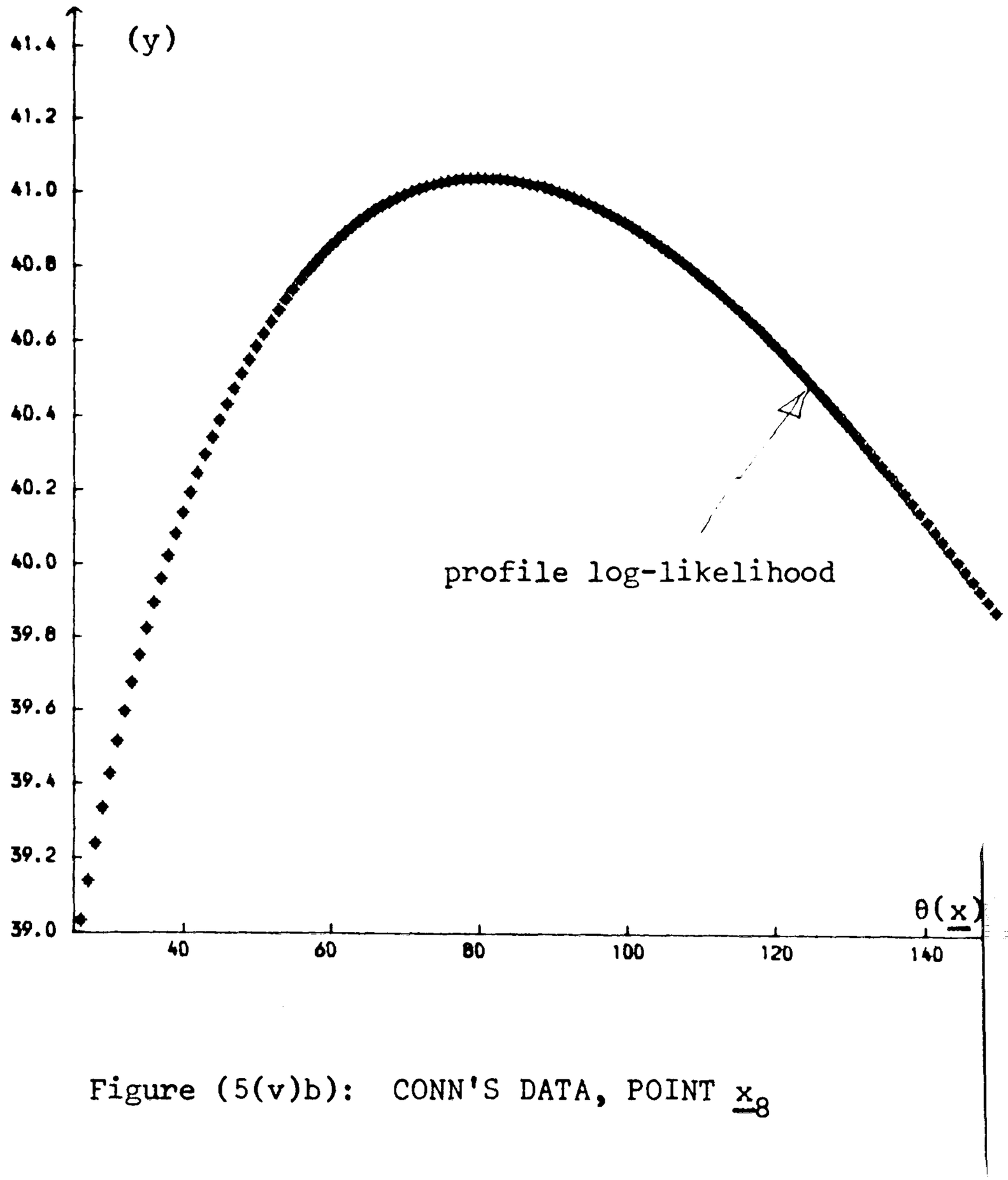


Figure (5(v)b): CONN'S DATA, POINT $\underline{x}_8$

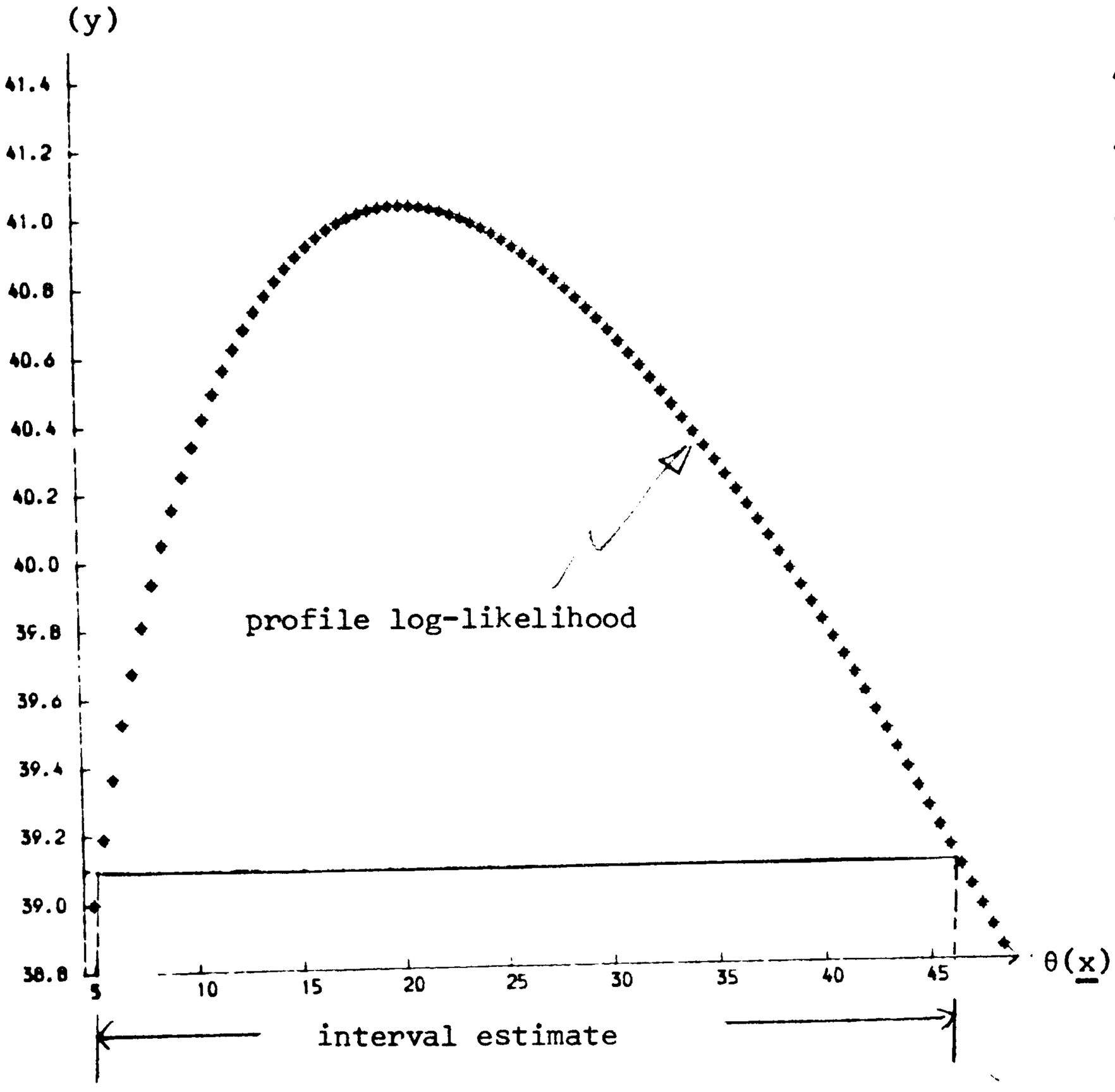


Figure [5(v)c] CONN'S DATA, POINT $\underline{x}_9$

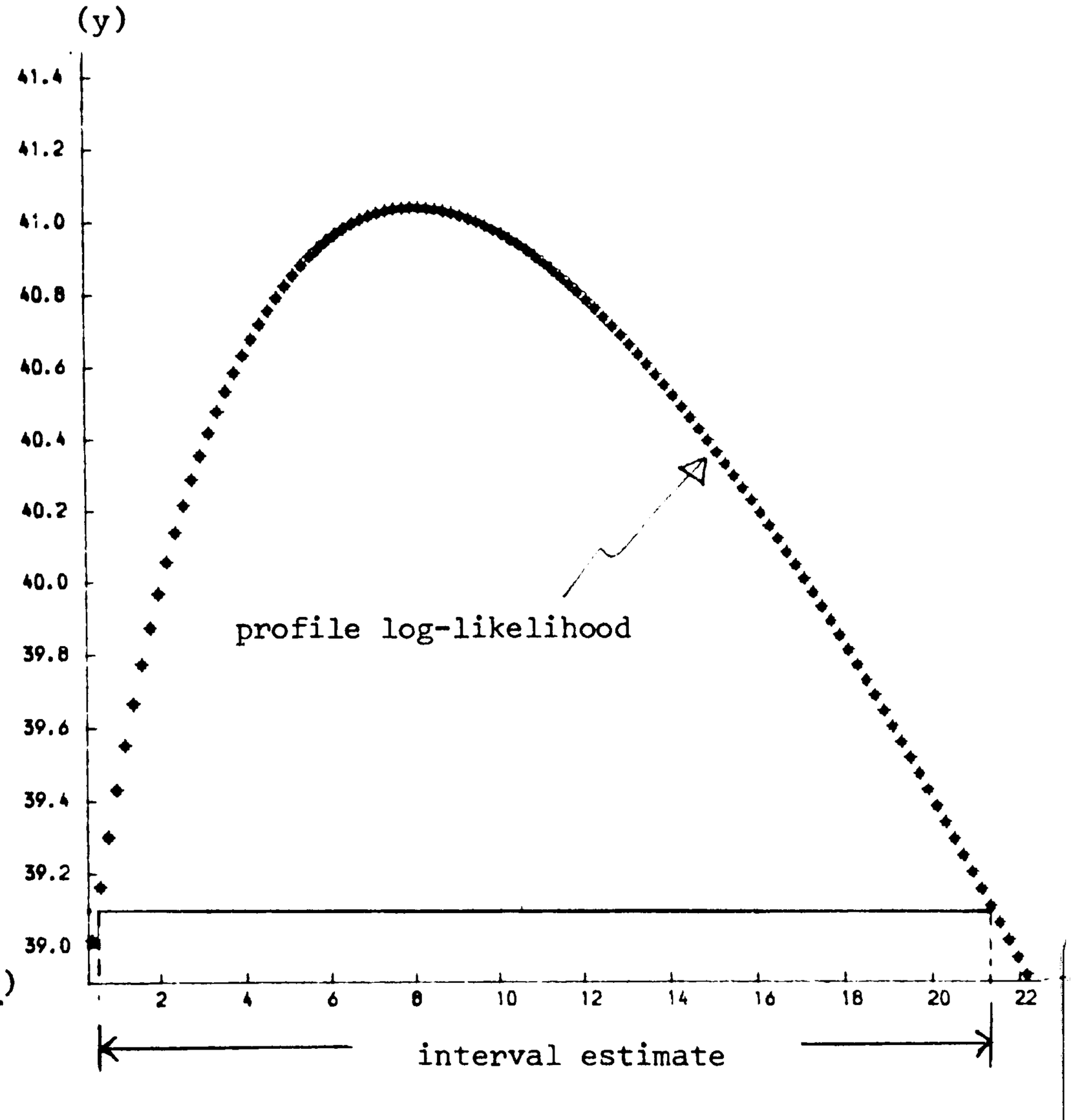


Figure [5(v)d] CONN'S DATA, POINT $\underline{x}_{18}$

CHAPTER 6

ROBUSTNESS TO NON-NORMALITY (p=2)

6.1: Introduction

In Chapters 2 and 3, we studied the distribution of the estimated log-odds, $\hat{e}(\underline{x})$, where the data in the training sets were normally distributed. In Chapters 4 and 5 we analysed some data sets, assuming that the data were normally distributed. We now propose to look at the effect of non-normality in the distribution of $\underline{x}$ on,

(i) the distribution of $\hat{\theta}$

(ii) the performance of approximate interval estimates for θ which are based on the assumption of normal distributions.

We will carry out a simulation study essentially continuing the study in Chapter 3. In particular, we will look at the simulations;

$$SA(1,2) = \left[\begin{array}{l} n_1 = 40 = n_2,\ p = 2 \\ \underline{\mu}^T = (2,0),\ \Delta = 2.0 \\ d_1 = 4 = d_2 \end{array}\right]$$

$$SA(2,2) = \left[\begin{array}{l} \text{as for } SA(1,2) \\ \text{but } d_1 = 1 = d_2 \end{array}\right]$$

where all the variables are as defined in Section (3.1.C). The $\underline{x}$-points are defined with respect to the normal distribution. They are \$9090, \$9038, the origin, $\underline{\mu}^T$ and \$CC (see Section (3.1.C)).

We will make use of the two-component univariate normal mixture to generate non-normal data. The mixture distribution itself is of no major significance. It is a convenient tool for constructing data with varying degrees of skewness and kurtosis. In the simulation study we will generate multivariate random vectors whose components are distributed as appropriate independent univariate normal mixtures.

As in Chapter 3 we will use θ_T to denote the true log-odds. The l^{th} population is denoted by Π_l (l=1,2).

Our present study will be comparable to SA(1,2) and SA(2,2) as simulated in Chapter 3 if we insist that,

$$\left.\begin{array}{l} \text{for } \Pi_1\text{: } E(\underline{x}) = \underline{\mu} \text{ and } Var(\underline{x}) = \text{diag}(d_1, d_2) \\ \qquad\qquad\qquad\qquad\qquad\qquad d_1 > 0,\ d_2 > 0 \\ \text{for } \Pi_2\text{: } E(\underline{x}) = \underline{0} \text{ and } Var(\underline{x}) = I_2. \end{array}\right\} \text{—(6.1.1)}$$

6.2 Some useful univariate results

We will use the same measures of univariate skewness (β) and kurtosis (γ) as defined in Chapter 3. They are,

$$\beta = m_3/(m_2\sqrt{m_2}) \text{ and } \gamma = m_4/(m_2^2)$$

where $m_k = E\{(x - E(x))^k\}$.

For normal distributions β=0 and γ=3.

To evaluate β and γ for the univariate normal mixture we use results from Johnson and Kotz (1970a) page 89.

Let $\mu_r' = E(X^r)$

and $f(x) = \lambda N(\xi_1, \sigma_1^2) + \lambda' N(\xi_2, \sigma_2^2)$, $\lambda + \lambda' = 1$,

that is, X is distributed as a mixture of two normal distributions $[N(\xi_i, \sigma_i^2);\ i=1,2]$ with mixing weight λ.

Then,
$$\left.\begin{array}{l} \mu_1' = \lambda\,\xi_1 + \lambda'\xi_2 \\ \mu_2' = \lambda(\xi_1^2 + \sigma_1^2) + \lambda'(\xi_2^2 + \sigma_2^2) \\ \mu_3' = \lambda(\xi_1^3 + 3\xi_1\sigma_1^2) + \lambda'(\xi_2^3 + 3\xi_2\sigma_2^2) \\ \mu_4' = \lambda(\xi_1^4 + 6\xi_1^2\sigma_1^2 + 3\sigma_1^4) + \lambda'(\xi_2^4 + 6\xi_2^2\sigma_2^2 + 3\sigma_2^4) \end{array}\right\} \text{—(6.2.1)}$$

It is well known that the two-component univariate normal mixture is symmetrical if

(i) $\lambda = \frac{1}{2}$ and $\sigma_1 = \sigma_2$

or (ii) $\xi_1 = \xi_2$

We will make use of this property to generate symmetrically

distributed data.

6.3: GENERATION OF DATA

(A) A special univariate normal mixture

$$f(x) = \lambda N(a,b) + \lambda' N(-a,bc) \qquad (6.3.1)$$

Let $\lambda = \frac{1}{2}$. Using (6.2.1) $E(x) = \frac{1}{2}a - \frac{1}{2}a = 0$

$$E(x^2) = Var(x) = \frac{1}{2}(a^2 + b) + \frac{1}{2}(a^2 + bc)$$

$$= a^2 + \frac{b}{2}(1+c)$$

$$E(x^3) = \frac{1}{2}(a^3 + 3ab) + \frac{1}{2}(-a^3 - 3abc)$$

$$= \frac{3ab}{2}[1-c]$$

$$E(x^4) = \frac{1}{2}(a^4 + 6a^2b + 3b^2) + \frac{1}{2}(a^4 + 6a^2bc + 3b^2c^2)$$

$$= a^4 + 3a^2b[1 + c] + \frac{3b^2}{2}[1 + c^2]$$

Since $E(x) = 0$, $\beta = \dfrac{E(x^3)}{[E(x^2)]^{3/2}}$ and $\gamma = \dfrac{E(x^4)}{[E(x^2)]^2}$

In the simulations we generate the data for Π_2 as outlined in figure (6(i)). The constants are chosen to satisfy (6.1.1).

For Π_1, we generate another set of (n_1 x 2) x-variables. Each x is multiplied by k_1, and then k_2 is added to the product. We note that β and γ are invariant to linear transformations of the type $x^* = xk_1 + k_2$. We can therefore choose appropriate values of k_1 and k_2 such that (6.1.1) is satisfied for Π_1.

For each population we calculate.

$$\underline{\bar{x}} = \frac{1}{n}\sum_{r=1}^{n} \underline{x}_r \text{ and } S = \sum_{r=1}^{n} (\underline{x}_r - \underline{\bar{x}})(\underline{x}_r - \underline{\bar{x}})^T$$

The only other change to the simulations in Chapter 3 for SA(1,2) and SA(2,2) is the calculation of the true log-odds, i.e.

$$\theta_T = \log\left[\frac{f(\underline{x}|\Pi_1)}{f(\underline{x}|\Pi_2)}\right].$$

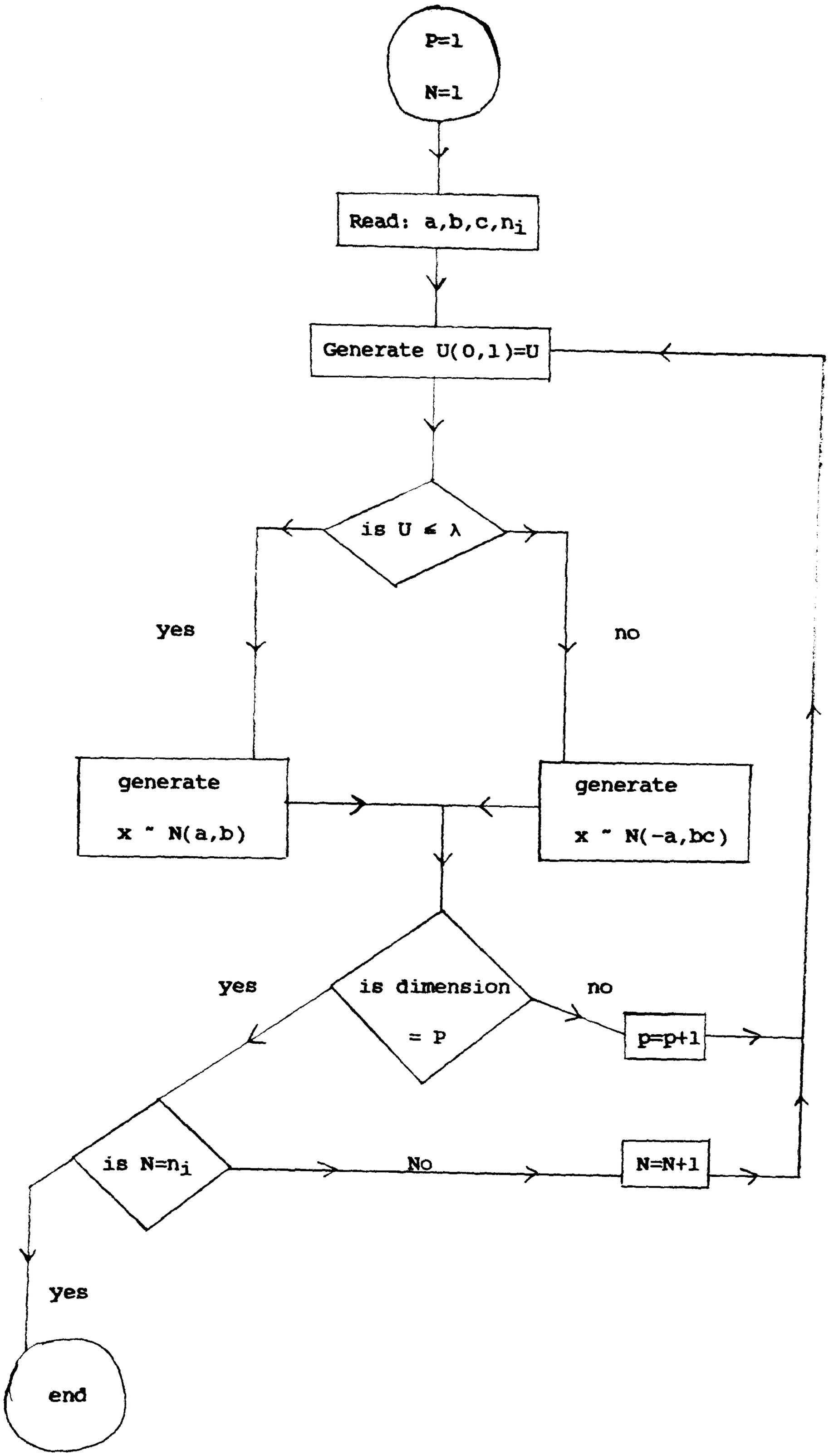


Figure (6(i)): Generation of data for π_2 (CASE (1))

Here, $f(\underline{x}|\Pi_i) = f(x_1|\Pi_i)\ f(x_2|\Pi_i)$, i=1,2, since x_1 and x_2 are independently generated, and $f(x_i|\Pi_i)$ is the density function of an appropriate univariate normal mixture.

To simplify our study we restrict ourselves to cases where β and γ are the same in each dimension and in each population. We now consider two particular forms of skewness and kurtosis and select values of a, b, c.

(B) Case (1): β=0 and increasing γ

POPULATION TWO:

Let a=0, thus $\beta = 0$.

By (6.1.1) we require Var(x) = 1, i.e. b(1+c) = 2. Clearly $\gamma = E(x^4) = 3b^2[1+c^2]/2$.

Alternatively:

Let c = 1.0, again $\beta = 0$ and we require $a^2 + b = 1$. Hence, $\gamma = E(x^4) = a^4 + 6a^2b + 3b^2$.

POPULATION ONE

By (6.1.1) we require $Var(\underline{x}) = k_1^2 I_2$ and $E(\underline{x}^T) = (\mu_1, \mu_2^T)$. Of course we use $k_1 = 1$ for SA(2,2) and k_1=2 for SA(1,2). Note $k_2 = \mu_1$ in the first dimension and $k_2 = \mu_2$ in the second.

(C) Case (2): Increasing skewness and fixed kurtosis (γ=3)

POPULATION TWO

We require $E(x^2) = 1$, i.e. $a^2 + b(1+c)/2 = 1$

$$\beta = E(x^3) = \frac{3ab}{2}(1-c)$$

$$\gamma = E(x^4) = a^4 + 3a^2b\,[1+c] + \frac{3b^2}{2}\,[1 + c^2]$$

These three equations contain three variables: a, b and c. By setting γ equal three and varying β we can obtain appropriate values of a, b, and c. It can be shown that for a suitable value of c using.

$$a^2 = [1 - \frac{b}{2}(1 + c)]$$

and $$b = \frac{-8(1 + c) \pm [64(1 + c)^2 + 32(1 - 10c + c^2)]^{1/2}}{2(1 - 10c + c^2)}$$

will ensure that Var(x) = 1 and $\gamma = 3$. Of course b must have positive values.

POPULATION ONE

The data are generated in the same manner as for population two with the appropriate transformations applied as before.

6.4: Simulation Study

(A) Statistics considered in simulations

For SA(1,2) we will only look at.

$\hat{CP2}$, $\hat{\epsilon 2}$, $\hat{V_S}(\hat{\theta_{NE}})$, $\hat{\beta 2}$, $\hat{\gamma 2}$

with definitions as given in sections (3.2.A) and (3.2.D).

For SA(2,2), in addition to the statistics considered in SA(1,2), we will also look at,

$\hat{CP1}$, $\hat{\epsilon 1}$, $\hat{V_S}(\hat{\theta_E})$, $\hat{\beta 1}$, $\hat{\gamma 1}$,

where again sections (3.2.A) and (3.2.D) give definitions.

Define,

(β, γ) ≡ (skewness, kurtosis) for the univariate normal mixtures that we generate,

$(\beta 1, \gamma 1)$ ≡ (skewness, kurtosis) for the empirical distribution of $\hat{\theta}(\underline{x})$ when we assume $\Omega_1 = \Omega_2$,

$(\beta 2, \gamma 2)$ ≡ (skewness, kurtosis) for the empirical distribution of $\hat{\theta}(\underline{x})$ when we assume $\Omega_1 \neq \Omega_2$.

(B) Choice of β and γ

We carried out some preliminary simulations first of all to identify interesting values of β and γ for the simulation study. Examples of these are given in Figures (6(iv)) and (6(v)) illustrating the effect of increasing skewness or kurtosis on CP2. Figures similar to figures (6(ii)) and (6(iii)) are useful for illustrating visually the

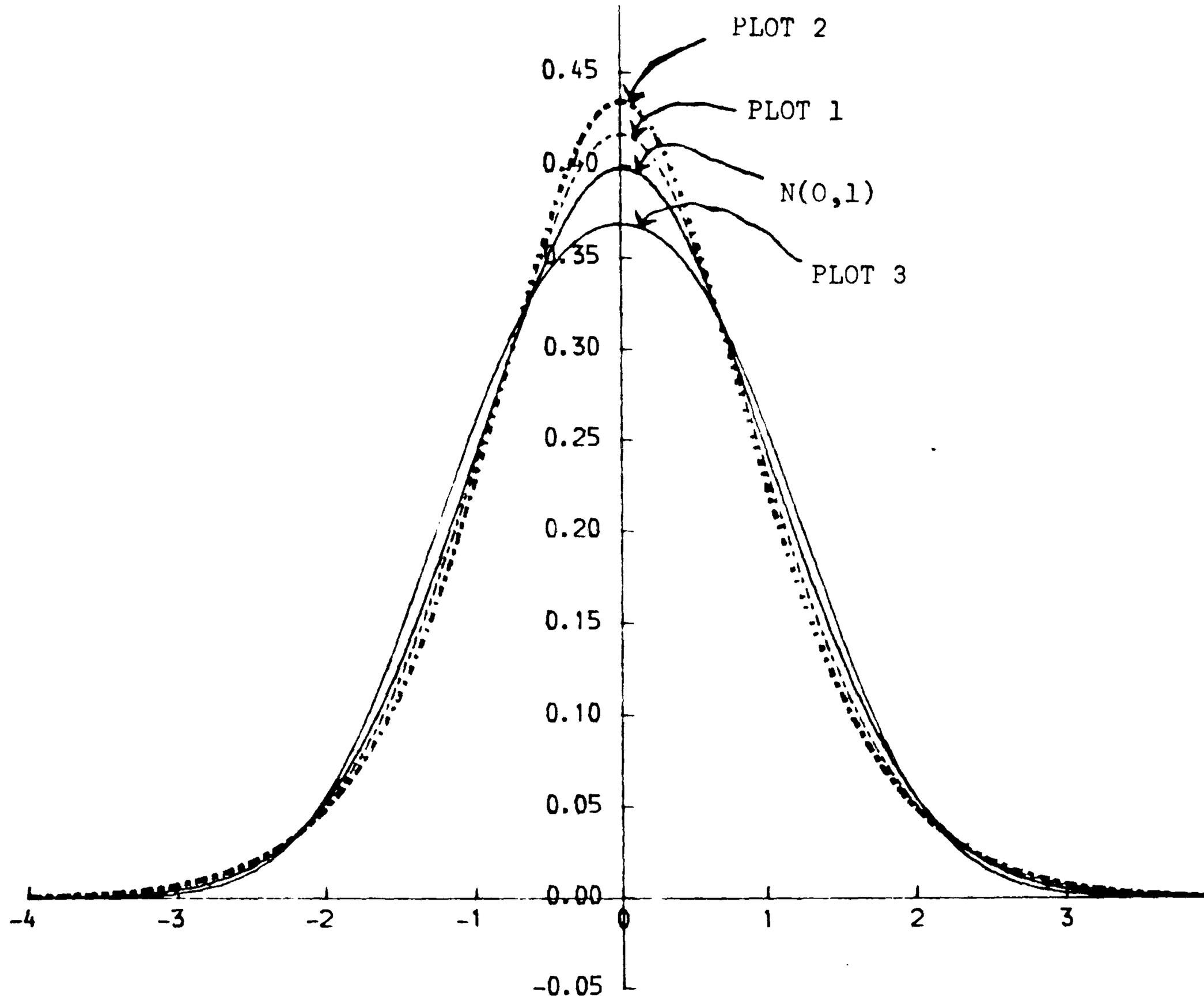


Figure (6(ii)): $\beta = 0$ and changing γ (case (1))

PLOT 1: A = 0.000 B = 0.667 C = 2.000 γ = 3.33

PLOT 2: A = 0.000 B = 1.450 C = 0.379 γ = 3.61

PLOT 3: A = 0.633 B = 0.600 C = 1.000 γ = 2.68

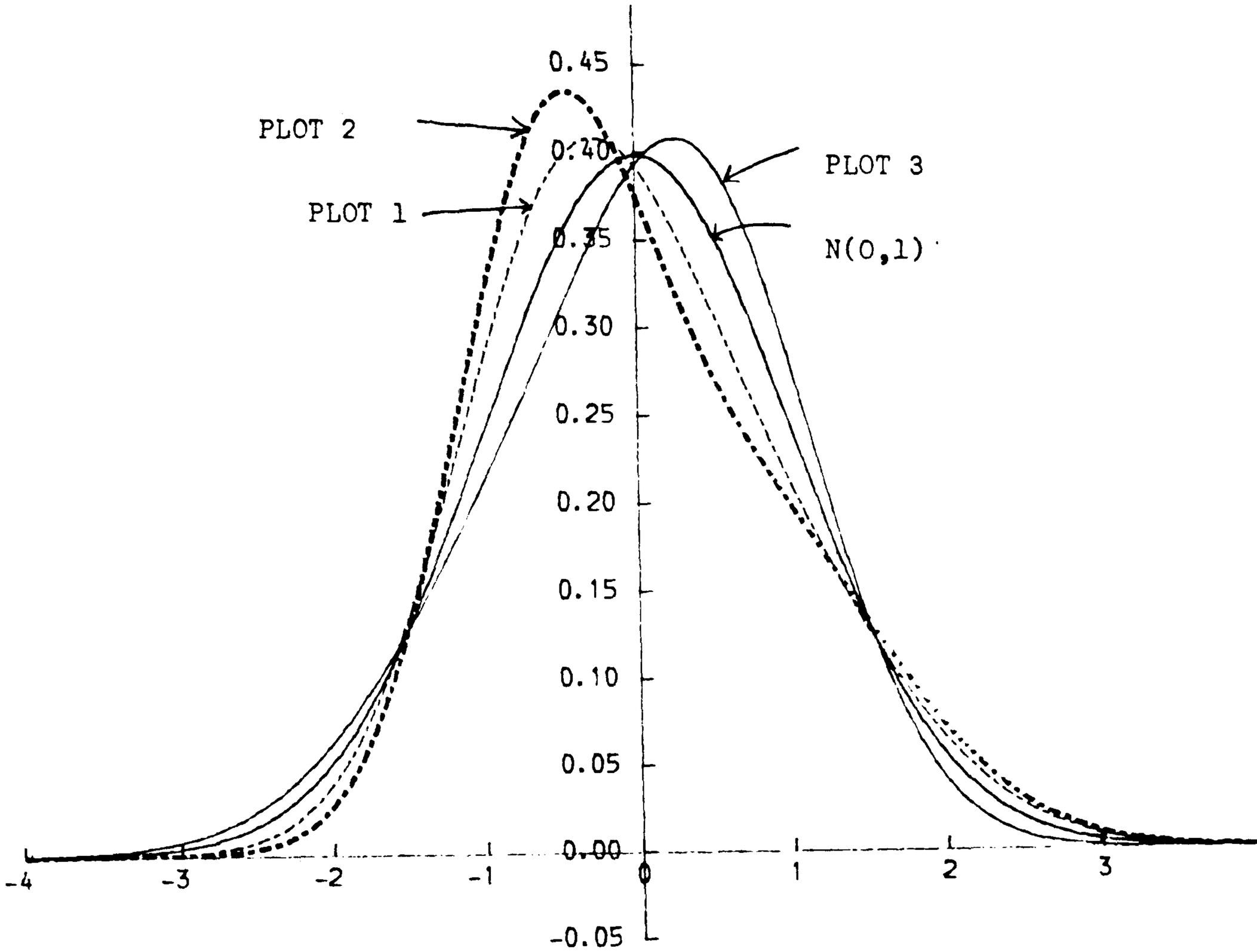


Figure (6(iii)): $\gamma = 3$ and changing β (case (2))

PLOT 1: A = 0.512 B = 0.952 C = 0.550 β = 0.33

PLOT 2: A = 0.589 B = 0.936 C = 0.395 β = 0.50

PLOT 3: A = 0.509 B = 0.529 C = 1.800 β =-0.32

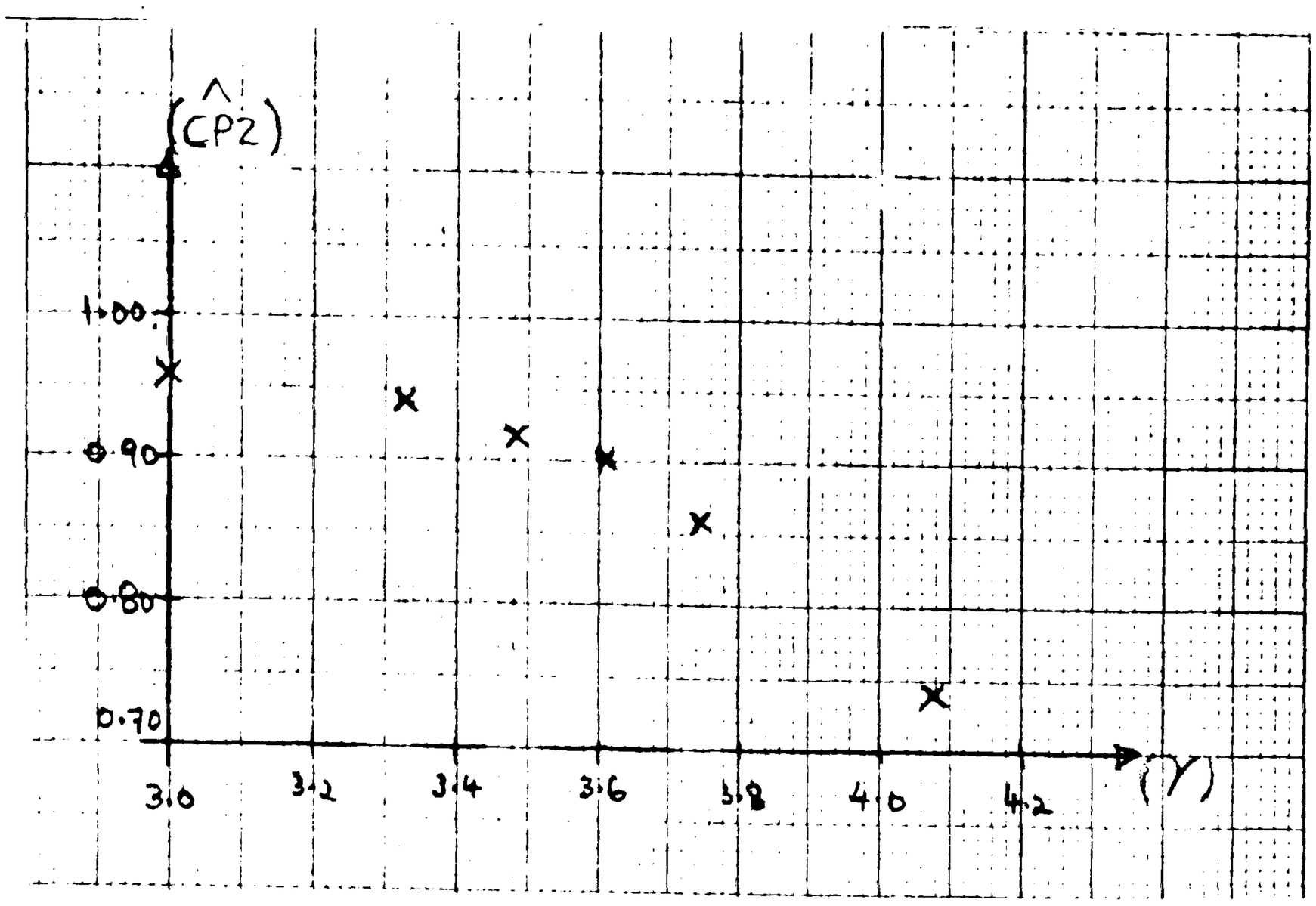


Figure (6(iv)): Effect of γ on $\hat{CP2}$ when

β = 0.0 and the x-point = origin

(case (1))

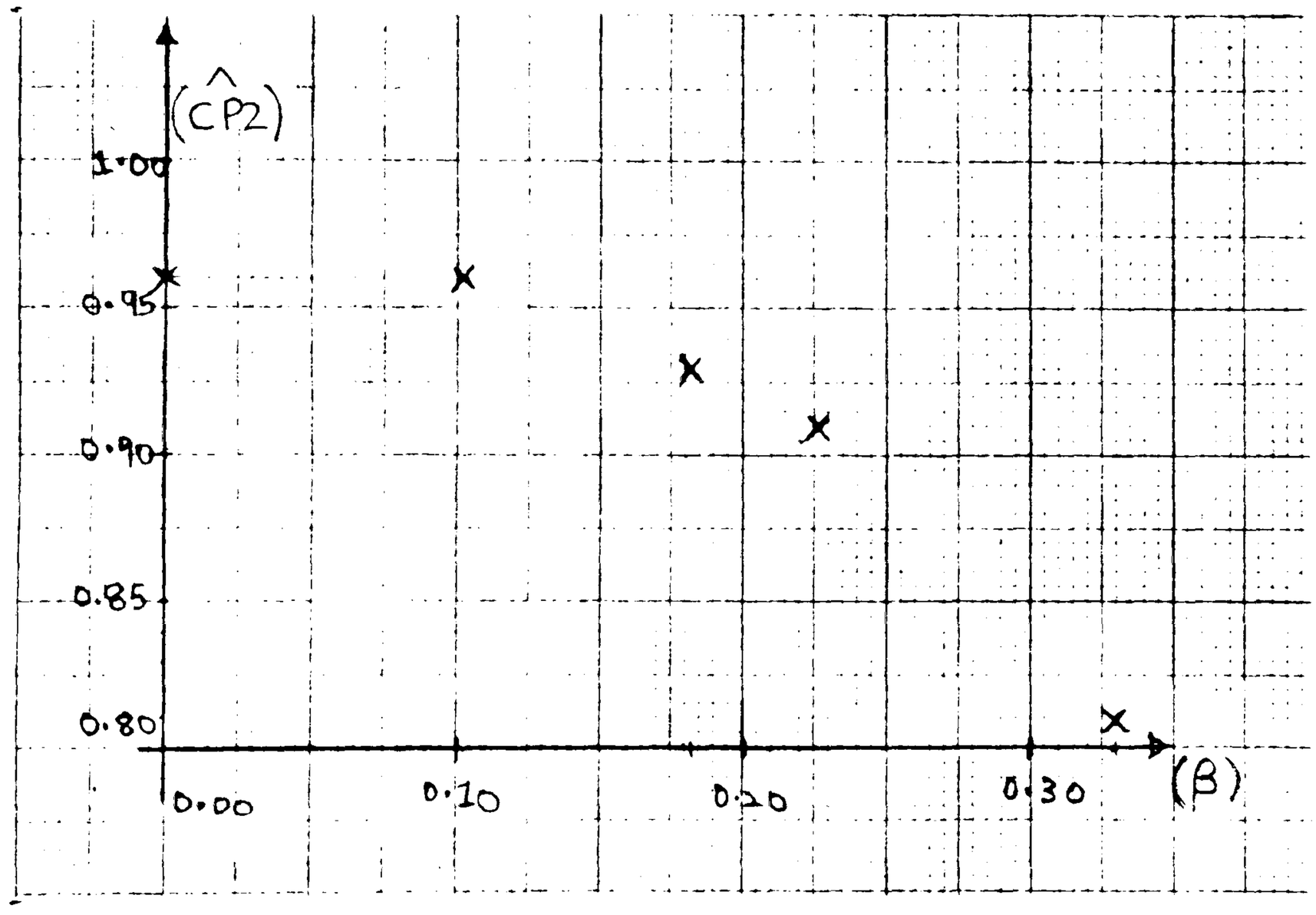


Figure (6(v)): Effect of β on $\hat{CP2}$ when

γ = 3.0 and the x-point = $CC

(case (2))

a	b	c	γ
0.00	$^2/_3$	2.0	$3^1/_3$
0.00	1.45	0.379310	3.61
0.632455	0.60	1.0	2.68

Table (6(i)): γ for case(1)

with $\beta=0$, $\lambda=^1/_2$, $k_1=2$

a	b	c	β
0.588983	0.936342	0.395	0.50
0.512155	0.951866	0.550	0.33
0.509135	0.529129	1.800	-0.32

Table (6(ii)): β for case(2)

with $\gamma=3$, $\lambda=^1/_2$, $k_1=2$

differences between the marginal densities of different distributions.

The values selected for β and γ (and hence a, b and c) are given in Tables (6(I)) and (6(II)). The corresponding marginal densities are illustrated in figures (6(II)) and (6(III)).

(C) Effect of changing β or γ in SA(1,2)

The results are given in Tables (6(iii)a), (6(iii)b) and (6(iii)c). All the statistics including $\hat{CP2}$ may vary with respect to β and γ, and the change in these statistics depends on the $\underline{x}$-point considered.

Generally poor values of $\hat{CP2}$ can be associated with larger values of $\hat{\epsilon 2}$. In Table (6(iii)c) values for $\hat{\beta 2}$ and $\hat{\gamma 2}$ are of the same order of magnitude (for varying β and γ) as for the normal data. However, there is no obvious systematic trend for $\hat{CP2}$ with respect to $\hat{\beta 2}$ and $\hat{\gamma 2}$. For instance, for the point $\underline{x}$ = \$CC, given $\gamma = 3.0$, the values of $\hat{\beta 2}$ and $\hat{\gamma 2}$ are very close to their normal distribution values, yet the corresponding $\hat{CP2}$ are the worst.

For the limited situations considered here, introducing skewness in the distribution of the data appears to have a potentially greater effect on $\hat{CP2}$ than the introduction of kurtosis, though clearly this requires further study.

(D) Effect of changing β or γ in SA(2,2)

The results are given in Tables (6(iv)a),....,(6(iv)e). Both when we assume $\Omega_1 = \Omega_2$ and $\Omega_1 \neq \Omega_2$ the results are similar to those described in section (6.4.C). Usually poor $\hat{CP1}$ and $\hat{CP2}$ values are associated with large biases. Relationships between $\hat{CP1}$ and $(\hat{\beta 1}, \hat{\gamma 1})$ or $\hat{CP2}$ and $(\hat{\beta 2}, \hat{\gamma 2})$ are less clear.

Generally $\hat{CP1}$ is smaller than $\hat{CP2}$ though often they are very close. Also $\hat{\epsilon 1}$ is always close to $\hat{\epsilon 2}$. The larger values for $\hat{\beta 2}$ and $\hat{\gamma 2}$ relative to $\hat{\beta 1}$ and $\hat{\gamma 1}$ are also present for the normally distributed data.

(6.5) Summary

Non-normality in the distribution of $\underline{x}$ does effect the confidence probabilities [CP1 and CP2] and the distribution of $\hat{\theta}_E(\underline{x})$ and $\hat{\theta}_{NE}(\underline{x})$.

The estimates of $\theta(\underline{x})$ tend to be biased when we vary β or γ. Increased bias can clearly affect the estimated confidence probabilities. This suggests that our interval estimates can be affected at least for some $\underline{x}$ values when the distribution of the data is non-normal. Possibly the levels of non-normality which we have introduced would be difficult to detect.

The choice of $\underline{x}$-point is critical with respect to,

(i) the effect of β, γ on the confidence probabilities and

(ii) the distribution of $\hat{\theta}(\underline{x})$.

Only a few combinations of (β, γ) were considered and we should emphasise the need for caution when interpreting the simulation results. However, estimation of the log-odds is based on the estimation of density functions and if we make incorrect assumptions about the underlying densities we must expect to get biased results.

x-point / (β,γ)	μ	origin	$CC	$9038	$9085
Normal data SA(1,2)	0.61 0.95	-1.89 0.96	-1.39 0.96	0.42 0.95	-0.98 0.98
β=0 $\gamma=3^{1}/_{3}$ (#)	0.72 0.91	-1.96 0.94	-1.39 0.95	0.56 0.91	-0.90 0.97
β=0 γ=3.61 (#)	0.80 0.89	-2.04 0.90	-1.39 0.94	0.68 0.88	-0.82 0.96
β=0 γ=2.68 (#)	0.52 0.97	-1.74 0.95	-1.39 0.97	0.27 0.97	-1.10 0.98
β=0.33 γ=3.0 (≠)	0.42 0.96	-1.72 0.93	-1.10 0.81	0.38 0.95	-1.19 0.98
β=0.50 γ=3.0 (≠)	0.28 0.95	-1.57 0.82	-0.86 0.47	0.32 0.95	-1.44 0.96
β=-0.32 γ=3.0 (≠)	0.87 0.90	-2.00 0.94	-1.67 0.82	0.46 0.95	-0.84 0.96

Table (6(iii)a): Each 'cell' contains

(i) θ_T (top number)

(ii) $\hat{CP2}$

(#) Note 1: see Table (6(i)) for corresponding values of (a,b,c)

(≠) Note 2: see Table (6(ii)) for corresponding vales of (a,b,c)

$\underline{x}$-point / (β,γ)	μ	origin	\$CC	\$9038	\$9085
Normal data	0.00	0.00	0.00	0.00	0.00
SA(1,2)	0.54	0.26	0.24	0.62	0.75
β=0	0.09	−0.08	0.00	0.13	0.08
$\gamma=3^1/_3$	0.57	0.27	0.25	0.63	0.78
β=0	0.16	−0.15	0.00	0.23	0.15
γ=3.61	0.59	0.28	0.26	0.64	0.80
β=0	−0.08	0.14	0.00	−0.13	−0.12
γ=2.68	0.51	0.25	0.23	0.60	0.75
β=0.33	−0.21	0.16	0.28	−0.06	−0.22
γ=3.0	0.59	0.26	0.24	0.66	0.73
β=0.50	−0.36	0.31	0.51	−0.13	−0.48
γ=3.0	0.63	0.26	0.24	0.69	0.71
β=−0.32	0.27	−0.11	−0.27	0.07	0.14
γ=3.0	0.48	0.26	0.24	0.57	0.81

Table (6(iii)b): Each cell contains

(i) $\hat{\varepsilon 2}$ (top number)

(ii) $\hat{V}_S(\hat{\theta}_{NE})$

$\underline{x}$-point / (β,γ)	μ	origin	\$CC	\$9038	\$9085
Normal data	0.93	-0.22	0.05	0.94	0.20
SA(1,2)	4.85	3.42	3.01	5.26	4.07
β=0	0.95	-0.17	0.02	0.88	0.20
$\gamma=3^1/_3$	5.07	3.27	3.09	4.70	4.84
β=0	0.98	-0.17	0.01	0.88	0.18
γ=3.61	5.44	3.41	3.05	4.66	4.52
β=0	0.91	-0.21	-0.01	0.88	0.12
γ=2.68	5.04	3.36	3.08	4.59	4.14
β=0.33	1.09	-0.09	0.03	0.99	0.21
γ=3.0	5.78	3.23	3.06	5.00	4.29
β=0.50	1.20	-0.05	0.07	1.07	0.29
γ=3.0	6.24	3.17	3.10	5.38	4.27
β=-0.32	0.84	-0.31	-0.05	0.82	0.03
γ=3.0	4.89	3.60	3.05	4.51	4.51

Table (6(iii)c): Each cell contains

(i) $\hat{\beta 2}$ (top number)

(ii) $\hat{\gamma 2}$

$\underline{x}$-point / (β,γ)	origin	\$CC	\$9038	\$9090
Normal data SA(2,2)	-2.00 0.94 0.95	0.00 0.96 0.96	1.81 0.95 0.95	0.00 0.96 0.98
β=0 $\gamma=3^1/_3$	-2.11 0.90 0.91	0.00 0.96 0.96	1.93 0.92 0.92	0.00 0.96 0.97
β=0 γ=3.61	-2.19 0.88 0.88	0.00 0.96 0.95	2.02 0.89 0.90	0.00 0.96 0.97
β=0 γ=2.68	-1.90 0.97 0.97	0.00 0.96 0.97	1.69 0.96 0.97	0.00 0.96 0.98
β=0.33 γ=3.0	-2.27 0.88 0.89	0.28 0.75 0.85	1.68 0.95 0.96	0.28 0.94 0.95
β=0.50 γ=3.0	-2.60 0.68 0.73	0.46 0.46 0.67	1.58 0.94 0.96	0.46 0.89 0.91
β=-0.32 γ=3.0	-1.81 0.94 0.96	-0.28 0.75 0.86	1.97 0.93 0.93	-0.28 0.94 0.97

Table (6(iv)a) Each cell contains

(i) θ_T (top number)

(ii) $\hat{CP1}$ (middle)

(iii) $\hat{CP2}$ (bottom)

$\underline{x}$-point / (β,γ)	origin	\$CC	\$9038	\$9090
Normal data	0.00	0.00	0.00	0.00
SA(2,2)	0.47	0.23	0.54	0.66
β=0	-0.10	0.00	0.11	0.00
$\gamma=3^1/_3$	0.49	0.23	0.55	0.66
β=0	-0.17	0.00	0.19	0.01
γ=3.61	0.50	0.23	0.57	0.67
β=0	0.09	0.00	-0.12	0.00
γ=2.68	0.44	0.23	0.53	0.65
β=0.33	-0.27	0.28	-0.14	0.28
γ=3.0	0.42	0.23	0.57	0.66
β=0.50	-0.61	0.44	-0.25	0.45
γ=3.0	0.40	0.23	0.59	0.66
β=-0.32	0.20	-0.27	0.17	-0.27
γ=3.0	0.50	0.23	0.51	0.66

Table (6(iv)b) Each cell contains

(i) $\hat{\epsilon}_1$ (top number)

(ii) $\hat{V}_S(\hat{\theta}_E)$

$\underline{x}$-point / (β,γ)	origin	$CC	$9038	$9090
Normal data	-0.71	0.00	0.57	0.00
SA(2,2)	3.95	3.68	3.77	3.46
β=0	-0.66	0.04	0.57	-0.01
$\gamma=3^1/_3$	3.66	3.65	3.80	3.54
β=0	-0.71	0.03	0.57	-0.06
γ=3.61	3.83	3.64	3.96	3.53
β=0	-0.65	0.03	0.57	-0.04
γ=2.68	3.77	3.41	3.98	3.37
β=0.33	-0.64	0.22	0.59	-0.03
γ=3.0	3.72	3.50	3.95	3.48
β=0.50	-0.63	0.31	0.63	-0.03
γ=3.0	3.69	3.57	4.09	3.47
β=-0.32	-0.68	-0.15	0.56	-0.06
γ=3.0	3.84	3.48	3.89	3.42

Table (6(iv)c) Each cell contains

(i) $\hat{\beta}_1$ (top number)

(ii) $\hat{\gamma}_1$

x-point / (β,γ)	origin	\$CC	\$9038	\$9090
Normal data	0.00	0.00	0.00	0.00
SA(2,2)	0.54	0.29	0.62	0.83
$\beta=0$	-0.09	0.00	0.10	0.00
$\gamma=3^1/_3$	0.57	0.30	0.63	0.85
$\beta=0$	-0.15	0.00	0.17	0.01
$\gamma=3.61$	0.59	0.30	0.65	0.87
$\beta=0$	0.08	0.00	-0.11	0.01
$\gamma=2.68$	0.51	0.28	0.60	0.80
$\beta=0.33$	-0.28	0.27	-0.16	0.28
$\gamma=3.0$	0.48	0.29	0.67	0.89
$\beta=0.50$	-0.62	0.43	-0.27	0.45
$\gamma=3.0$	0.44	0.29	0.71	0.93
$\beta=-0.32$	0.21	-0.26	0.17	-0.25
$\gamma=3.0$	0.60	0.29	0.57	0.76

Table (6(iv)d) Each cell contains

(i) $\hat{\epsilon}2$ (top number)

(ii) $\hat{V}_S(\hat{\theta}_{NE})$

x-point / (β,γ)	origin	\$CC	\$9038	\$9090
Normal data	-0.94	0.00	0.94	0.00
SA(2,2)	4.93	3.38	5.14	4.26
β=0	-0.97	0.03	0.88	-0.04
$\gamma=3^1/_3$	4.94	3.34	4.76	4.45
β=0	-0.94	0.01	0.93	-0.14
γ=3.61	4.90	3.17	5.12	4.25
β=0	-1.05	0.01	0.93	-0.08
γ=2.68	6.26	3.27	5.12	3.97
β=0.33	-0.91	0.16	1.01	-0.09
γ=3.0	5.79	3.36	5.31	4.25
β=0.50	-0.82	0.26	1.10	-0.09
γ=3.0	5.40	3.60	5.74	4.59
β=-0.32	-1.18	-0.14	0.86	-0.12
γ=3.0	6.74	3.29	4.87	3.98

Table (6(iv)e) Each cell contains

(i) $\hat{\beta}2$ (top number)

(ii) $\hat{\gamma}2$

CHAPTER 7

SHORTCOMINGS AND FURTHER WORK

(7.1) Shortcomings

Research similar to some of our work in Chapters two and three has been done in Holland. A summary of this work is given in Ambergen and Schaafsma (1984). The exact variance of $\hat{\theta}(\underline{x})$ when $\Omega_1 = \Omega_2$ is available in Schaafsma and Van Vark (1977) for the univariate case. The asymptotic variance of $\hat{\theta}(\underline{x})$ when $\Omega_1 = \Omega_2$ is given in Schaafsma and Van Vark (1979) for the multivariate case. For the unequal covariance case, Ambergen and Schaafsma (1984) give the asymptotic variance of $\hat{\theta}(\underline{x})$, with details of the mathematics in Ambergen and Schaafsma (1983). One stage of the calculations in Ambergen and Schaafsma (1983) uses asymptotic results, i.e. Fishers information matrix, to derive the covariance matrix of the distinct elements of a Wishart Matrix. This is in contrast to our approach using the exact covariance matrix. We emphasise that the work in this thesis was carried out independently of this Dutch work and the above papers only came to our attention recently.

When studying the problem of interval estimation of $\hat{\theta}(\underline{x})$ in small sample situations the techniques of bootstrap and profile likelihood have not been fully investigated. For the bootstrap method other interval estimates could be derived using for example the "bias-corrected percentile" method. For the profile-likelihood method we have not given a rigourous proof of the generality of the success of the Lagranglan method. Only one data-set was analysed using these methods.

Another formula for the approximate variance of $\hat{\theta}(\underline{x})$ when $\Omega_1 \neq \Omega_2$, i.e. $BV(\hat{\theta}_{NE}(\underline{x}))$ given in Section (2.4.C), was derived only at the end of the simulation study in Chapter 3. We have not used $BV(\hat{\theta}_{NE}(\underline{x}))$ in these simulations or anywhere else.

When $\Omega_1 = \Omega_2$, $\text{var}(\hat{\theta}(\underline{x}))$ has the form $a\theta(\underline{x})^2 + b\theta(\underline{x}) + g(\phi(\underline{x}), \Delta)$. One approximate 95% interval estimate is the set of $\theta(\underline{x})$ that satisfies:

$$(\theta(\underline{x}) - \hat{\theta}(\underline{x}))^2 < 1.96^2 \, [a\theta(\underline{x})^2 + b\theta(\underline{x}) + g(\hat{\phi}(\underline{x}), \hat{\Delta})]$$

This interval estimate was shown (Critchley and Ford (1985)) to give superior confidence probabilities when compared to using $\hat{\theta}(\underline{x}) \pm 1.96\sqrt{\text{Var}(\hat{\theta}(\underline{x}))}$. Similar ideas were not tried out for the unequal covariance case.

The test for the equality of covariance matrices, in Chapter 3, was shown to be insensitive to small deviations from equality of the covariance matrices. Clearly a more powerful test of $\Omega_1 = \Omega_2$ is required. Possibly using an alternative $\Omega_1 = d\Omega_2$ would give increased power against some alternatives.

We do not have expressions for the third and fourth moments of $\hat{\theta}(\underline{x})$. If available, these moments could be used to provide a better approximation to the distribution of $\hat{\theta}(\underline{x})$ using a suitable four parameter family.

(7.2) Further work

Our study could extend further into the techniques of logistic discrimination (Anderson (1972)). The logistic form for the posterior probabilities are:

$$\text{Prob}(\Pi_1|\underline{x}) = [\exp(\underline{a}^T\underline{x})]. \quad \text{Prob}(\Pi_2|\underline{x})$$

$$\text{Prob}(\Pi_2|\underline{x}) = 1/(1 + \exp(\underline{a}^T\underline{x})) \qquad (7.1)$$

where $\underline{a}$ is the vector of unknown parameters.

A particular advantage in using logistic discrimination is that (7.1) is satisfied by several distributions for the underlying distribution of $\underline{x}$ for each population other than the multivariate normal.

It may be possible to assess the information loss in using logistic discrimination rather than the 'linear discriminant' approach

when the data really comes from multivariate normal distributions with $\Omega_1 = \Omega_2$. This could be done by comparing $\mathrm{Var}(\hat{\theta}_E(\underline{x}))$ with $\mathrm{Var}(\hat{\theta}_L(\underline{x}))$ where $\hat{\theta}_L(\underline{x})$ is the estimated log-odds under the logistic assumption. Clearly only asymptotic formulae would be available for $\mathrm{Var}(\hat{\theta}_L(\underline{x}))$.

Appendix 2.1

To evaluate $\sum_{ijrs} (\Omega - \underline{b}\,\underline{b}^T)_{ij}\,(\Omega - \underline{b}\,\underline{b}^T)_{rs}\,\mathrm{cov}(s^{ij}, s^{rs})$

We look at the four components separately. Note that

$\Omega = \{\omega_{ij}\}$ and $a\Omega^{-1} = \{\delta^{ij}\}$. We use 'tr' to denote trace.

$$(1)\quad \sum_i \sum_j \sum_r \sum_s \omega_{ij}\omega_{rs}[2\delta^{ij}\delta^{rs} + (t-2)(\delta^{ir}\delta^{js} + \delta^{is}\delta^{jr})]$$

$$2 \sum_i \sum_j \omega_{ij}\delta^{ji} \sum_r \sum_s \omega_{rs}\delta^{sr} + (t-2) \sum_i \sum_j \sum_r \sum_s \omega_{ij}\delta^{js}\omega_{sr}\delta^{ri}$$

$$+ (t-2) \sum_i \sum_j \sum_r \sum_s \omega_{ij}\delta^{jr}\omega_{rs}\delta^{si}$$

$$= 2\,\mathrm{tr}[\Omega(a\Omega^{-1})]\,\mathrm{tr}[\Omega(a\Omega^{-1})] + (t-2)\,\mathrm{tr}\,[\Omega a\Omega^{-1}\Omega a\Omega^{-1}]$$

$$+ (t-2)\,\mathrm{tr}\,[\Omega a\Omega^{-1}\Omega a\Omega^{-1}]$$

$$= 2p^2a^2 + 2(t-2)pa^2$$

$$= 2pa^2[n-2]$$

$$(2)\quad \sum_i \sum_j \sum_r \sum_s b_ib_jb_rb_s\,[2\delta^{ij}\delta^{rs} + (t-2)\{\delta^{ir}\delta^{js} + \delta^{is}\delta^{jr}\}]$$

$$= 2 \sum_i \sum_j b_i\,\delta^{ij}b_j \sum_r \sum_s b_r\,\delta^{rs}b_s$$

$$+ (t-2)\,\{\sum_i \sum_r b_ib_r\,\delta^{ir} \sum_j \sum_s b_jb_s\,\delta^{js}\}$$

$$+ (t-2)\,\{\sum_i \sum_s b_ib_s\,\delta^{is} \sum_j \sum_r b_jb_r\,\delta^{jr}\}$$

$$= 2(\underline{b}^T(a\Omega^{-1})\,\underline{b})^2 + 2(t-2)\,(\underline{b}^T(a\Omega^{-1})\,\underline{b})^2$$

$$= a^2\,(2t-2)\,[\underline{b}^T\Omega^{-1}\underline{b}]^2$$

$$(3)\quad \sum_i \sum_j \sum_r \sum_s b_ib_j\,\omega_{rs}\,[2\delta^{ij}\delta^{rs} + (t-2)\,\{\delta^{ir}\delta^{js} + \delta^{is}\delta^{jr}\}]$$

$$= 2 \sum_i \sum_j b_ib_j\,\delta^{ij} \sum_r \sum_s \omega_{rs}\delta^{rs} + \sum_{ijrs} (t-2)b_i\delta^{ir}\omega_{rs}\delta^{sj}b_j$$

$$+ (t-2) \sum_{ijrs} b_i\delta^{is}\omega_{sr}\delta^{rj}b_j$$

$$= 2\,\underline{b}^T(a\Omega^{-1})\underline{b}.\mathrm{tr}[\Omega(a\Omega^{-1})] + (t-2)\{\underline{b}^T(a\Omega^{-1})\Omega(a\Omega^{-1})\underline{b}\}$$

$$+ (t-2)\,\underline{b}^T\,(a\Omega^{-1})\Omega(a\Omega^{-1})\underline{b}$$

$$= 2a^2\,(\underline{b}^T\Omega^{-1}\underline{b})(n-2)$$

$$(4)\quad \sum_{ijrs} \omega_{ij}\,b_rb_s\,[2\delta^{ij}\delta^{rs} + (t-2)\,\{\delta^{ir}\delta^{js} + \delta^{is}\delta^{jr}\}]$$

$$= \text{as in (3).}$$

Appendix 2.2

We quote some technical results here.

Let S be a k x k symmetric matrix of independent real variables. Denote S^{-1} by (s^{ij}); and $(0,0,\ldots,\underset{\substack{\uparrow \\ g^{th}\text{ position}}}{1},\ldots,0)$ by $\underline{e}_g^T$

RESULT (1)

$$\frac{\delta S}{\delta s^{gh}} = \begin{cases} -(S\underline{e}_g\underline{e}_h^T S + S\underline{e}_h\underline{e}_g^T S) & \text{if } g \neq h \\ -(S\underline{e}_g\underline{e}_g^T S) & \text{if } g = h \end{cases}$$

The result is given in Corollary (10.8.10) of Graybill (1983), page 358. The case g=h is proven in the same book and given as Theorem (10.8.10). Clearly,

$$\frac{\delta s_{kl}}{\delta s^{gh}} = -\frac{1}{2}(2-\delta_{gh})(s_{kg}s_{hl} + s_{kh}s_{gl})$$

$$\text{where } \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise.} \end{cases}$$

RESULT (2)

[Mardia, Kent, Bibby (1979), Problem 3.4.17(c), pg 92]

Let $\Omega = (\sigma_{ij})$

Then, if $M \sim W_p(n,\Omega)$

$$\text{Var}(m_{ij}) = n(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})$$

$$\text{cov}(m_{ij},m_{kl}) = n(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$$

RESULT (3)

[Graybill (1983), Theorem (10.8.8)]

$$\frac{\delta(\log|S^{-1}|)}{\delta s^{gh}} = (2-\delta_{gh})s_{gh}$$

where $S^{-1} = (s^{ij})$

and $\delta_{gh} = \begin{cases} 1 \text{ if } g=h \\ 0 \text{ otherwise.} \end{cases}$

Appendix (3.1)

We have for population Π_i, $\underline{x} \sim Np(\underline{\mu}_i, \Omega_i)$, i=1,2. We apply a linear transformation $\underline{x} \rightarrow A\underline{x} + \underline{b} = \underline{x}^*$ and convert the distributions into canonical form, viz:-

Π_1: $\underline{x}^* \sim Np(\underline{\mu}, D)$ and Π_2: $\underline{x}^* \sim Np(\underline{0}, I)$

where D = diagonal matrix

I = identity matrix

$\underline{0} = (0,\ldots,0)^T$ and $\underline{\mu}$ = a p-variate vector.

Proof: Require $A\,\Omega_2\,A^T \rightarrow I$, $A\,\Omega_1\,A^T \rightarrow D$

Since Ω_2 is symmetric and non-singular, we can write it in spectral form;

$\Omega_2 = Q_2\,\Lambda_2\,Q_2^T$, Q_2 = matrix of eigenvectors

Λ_2 = matrix of eigenvalues.

Let $P = \Lambda_2^{-1/2}\,Q_2^T$, then $P\,\Omega_2\,P^T = I$

with P non-singular.

Let $C = P\,\Omega_1\,P^T$, clearly C is symmetric and has it's own spectral decomposition, say

$$C = Q_C\,\Lambda_C\,Q_C^T$$

or $Q_C^T\,C\,Q_C = \Lambda_C$

Finally let $A = Q_C^T P$

$$\text{Thus } A\Omega_2A^T = Q_C^T P\Omega_2 P^T Q_C = Q_C^T I Q_C = I$$

$$\text{and } A\Omega_1A^T = Q_C^T P\Omega_1 P^T Q_C = Q_C^T C Q_C = \Lambda_C.$$

The means in the $\underline{x}^*$ space can be obtained as follows:

$$\underline{\mu}_1^* = A\underline{\mu}_1 + \underline{b} = \underline{\mu}$$

$$\underline{\mu}_2^* = A\underline{\mu}_2 + \underline{b} = \underline{0}$$

Hence $\underline{b} = -A\underline{\mu}_2$

and $\underline{\mu} = A(\underline{\mu}_1 - \underline{\mu}_2)$

Appendix (3.2)

Aim: Generating (i) $\bar{\underline{y}} \sim Np\{\underline{\mu}, \frac{1}{L}D\}$; a p-dimensional normal distribution.

(ii) $S \sim Wp\{n, D\}$; a p-variate wishart distribution with n degrees of freedom and scale matrix D

where $D = diag\{d_1,\ldots,d_p\}$

Method

We made use of two random number generators from the NAG (1978) computer package, which are

(a) G05DDF(0,1): generates a univariate N(0,1)

(b) G05DHF(N,I): generates a univariate chi-square with N degrees of freedom (I=constant).

Generating the mean vector $\bar{\underline{y}}$ is equivalent to generating the elements of the vector independently, since the covariance matrix, D, is diagonal.

To generate the Wishart matrix S, we made use of Bartlett's decomposition (description by Kendall and Stuart (1966) Vol.3, p262). We need to generate the triangular matrix B, where,

$$B = \begin{bmatrix} (\underline{z}_1^T\underline{z}_1)^{1/2} & 0 & 0 & \ldots & 0 \\ b_{21} & (\underline{z}_2^T\underline{z}_2)^{1/2} & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ b_{p1} & b_{p2} & & \cdot & (\underline{z}_p^T\underline{z}_p)^{1/2} \end{bmatrix}$$

The elements of B are mutually independent and

$b_{jk} \sim N(0,1)$, $k=1,\ldots,(p-1)$

$\underline{z}_k^T\underline{z}_k \sim \chi^2(n-k+1)$.

So we generate matrix B by independently generating its elements. It can be shown that $A = BB^T \sim Wp\{n, I\}$

I = identity matrix

and finally, matrix S is obtained since

$S = D^{1/2} A D^{1/2} \sim Wp\{n, D\}$.

Appendix (3.3)

If $\hat{\theta}_{NE}$ is approximately Normally distributed, then

$$\hat{M}(2) \doteq V(\hat{\theta}_{NE})\, \chi^2\ (NREPL-1)$$

$$\doteq V(\hat{\theta}_{NE})\, N(NREPL-1,\ 2(NREPL-1))$$

since NREPL is large.

Therefore an approximate confidence interval for $V(\hat{\theta}_{NE})$ would be obtained from

$$\frac{\hat{M}(2)}{(NREPL-1) + 1.96\sqrt{VAR}}, \quad \frac{\hat{M}(2)}{(NREPL-1) - 1.96\sqrt{VAR}}$$

where VAR = 2(NREPL-1)

Since NREPL = 10,000, the approximate 95% Confidence Interval is

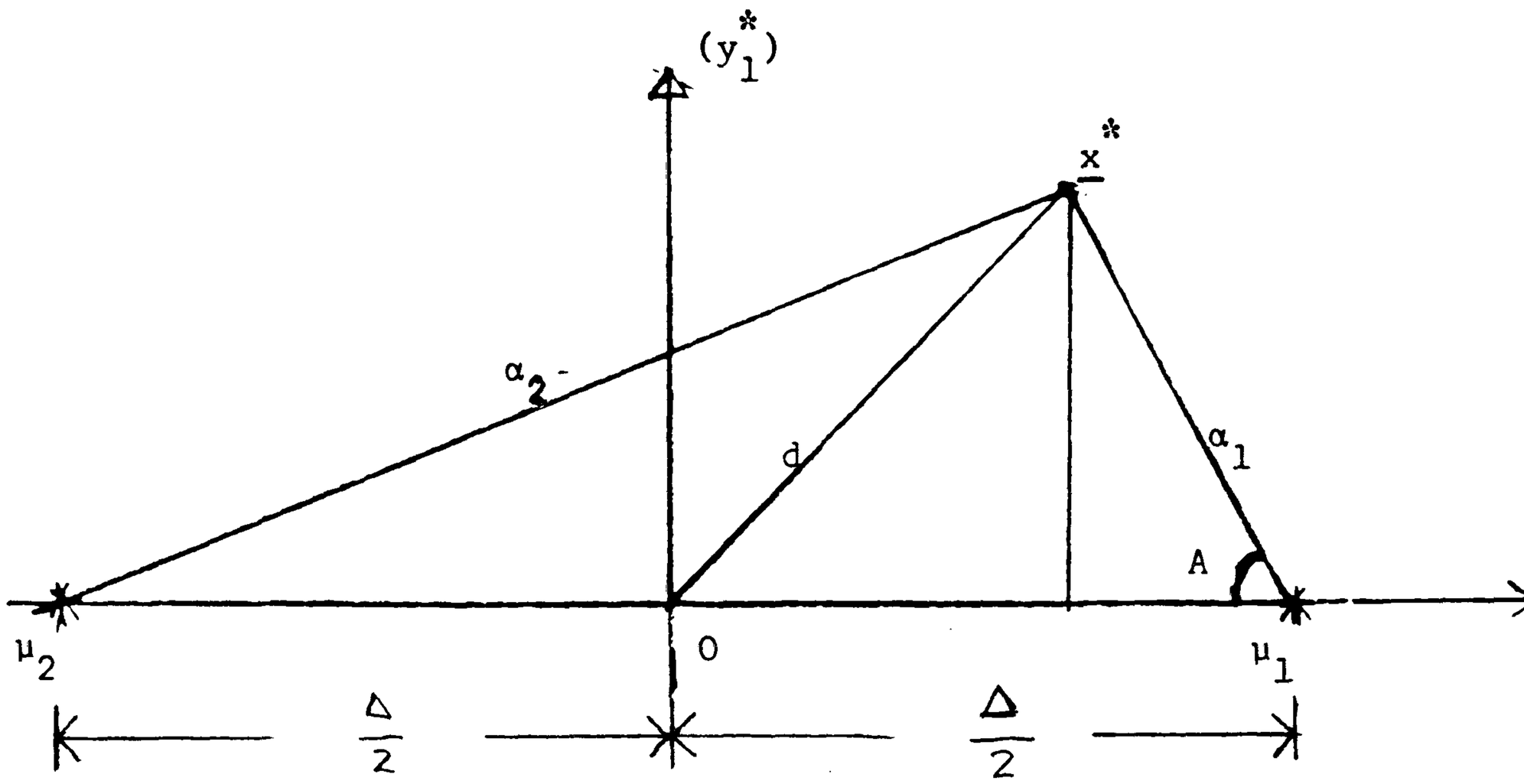
$$[0.97\ \hat{V}_S(\hat{\theta}_{NE}),\ 1.03\ \hat{V}_S(\hat{\theta}_{NE})]$$

In practice of course for small n_1, n_2, $\hat{\theta}_{NE}$ will not be exactly normally distributed. However, the above formula should still provide a useful rule of thumb.

Appendix (4.1)

When $\Omega_1 = \Omega_2$ each $\underline{x}$-point in p dimensional space is defined by $(\Delta^2,\ \alpha_1(\underline{x}),\ \alpha_2(\underline{x}))$.

We can therefore define our $(x^*,\ y^*)$ plane, illustrated by the following diagram.



Let d = distance between origin and $\underline{x}^*$. Using the cosine rule formula on angle A, and the fact that

$$d^2 = (x_1^*)^2 + (y_1^*)^2$$

$$\text{and} \quad \theta = \frac{1}{2}(\alpha_2^2 - \alpha_1^2), \quad \text{where } \alpha_i^2 = \alpha_i^2(\underline{x}^*)$$

we can show that $x_1^* = \theta/\Delta$

$$\text{and } y^* = \left[\phi - \frac{\Delta^2}{4} - \frac{\theta^2}{\Delta^2}\right]^{1/2}$$

where $\phi = \frac{1}{2}(\alpha_1^2 + \alpha_2^2)$ [see also Critchley and Ford (1984)].

APPENDIX (4.2.A) IRIS

IRIS VERSICOLOUR				IRIS VIRGINICA				
SL	SW	PL	PW	SL	SW	PL	PW	
65.	28.	46.	15.	64.	28.	56.	22.	
62.	22.	45.	15.	67.	31.	56.	24.	SL=SEPAL LENGTH
59.	32.	48.	18.	63.	28.	51.	15.	SW=SEPAL WIDTH
61.	30.	46.	14.	69.	31.	51.	23.	PL=PETAL LENGTH
60.	27.	51.	16.	65.	30.	52.	20.	PW=PETAL WIDTH
56.	25.	39.	11.	65.	30.	55.	18.	
57.	28.	45.	13.	58.	27.	51.	19.	
63.	33.	47.	16.	68.	32.	59.	23.	N1=50=N2
70.	32.	47.	14.	62.	34.	54.	23.	
64.	32.	45.	15.	77.	38.	67.	22.	
61.	28.	40.	13.	67.	33.	57.	25.	
55.	24.	38.	11.	76.	30.	66.	21.	
54.	30.	45.	15.	49.	25.	45.	17.	
58.	26.	40.	12.	67.	30.	52.	23.	
55.	26.	44.	12.	59.	30.	51.	18.	
50.	23.	33.	10.	63.	25.	50.	19.	
67.	31.	44.	14.	64.	32.	53.	23.	
56.	30.	45.	15.	79.	38.	64.	20.	
58.	27.	41.	10.	67.	33.	57.	21.	
60.	29.	45.	15.	77.	28.	67.	20.	
57.	26.	35.	10.	63.	27.	49.	18.	
57.	29.	42.	13.	72.	32.	60.	18.	
49.	24.	33.	10.	61.	30.	49.	18.	
56.	27.	42.	13.	61.	26.	56.	14.	
57.	30.	42.	12.	64.	28.	56.	21.	
66.	29.	46.	13.	62.	28.	48.	18.	
52.	27.	39.	14.	77.	30.	61.	23.	
60.	34.	45.	16.	63.	34.	56.	24.	
50.	20.	35.	10.	58.	27.	51.	19.	
55.	24.	37.	10.	72.	30.	58.	16.	
58.	27.	39.	12.	71.	30.	59.	21.	
62.	29.	43.	13.	64.	31.	55.	18.	
59.	30.	42.	15.	60.	30.	48.	18.	
60.	22.	40.	10.	63.	29.	56.	18.	
67.	31.	47.	15.	77.	26.	69.	23.	
63.	23.	44.	13.	60.	22.	50.	15.	
56.	30.	41.	13.	69.	32.	57.	23.	
63.	25.	49.	15.	74.	28.	61.	19.	
61.	28.	47.	12.	56.	28.	49.	20.	
64.	29.	43.	13.	73.	29.	63.	18.	
51.	25.	30.	11.	67.	25.	58.	18.	
57.	28.	41.	13.	65.	30.	58.	22.	
61.	29.	47.	14.	69.	31.	54.	21.	
56.	29.	36.	13.	72.	36.	61.	25.	
69.	31.	49.	15.	65.	32.	51.	20.	
55.	25.	40.	13.	64.	27.	53.	19.	
55.	23.	40.	13.	68.	30.	55.	21.	
66.	30.	44.	14.	57.	25.	50.	20.	
68.	28.	48.	14.	58.	28.	51.	24.	
67.	30.	50.	17.	63.	33	60.	25.	

APPENDIX (4.2.B) CONNS SYNDROME

	Z1	Z2	Z3	Z4
1 00000	3.68888	3.13549	3.41115	3.82864
1.00000	3.61092	3.43399	3.29353	3.80666
1.00000	3.52636	3.40120	3.29584	1.94591
1.00000	3.87120	3.33220	3.49651	3.49651
1 00000	3.71357	3.58352	3.18221	3.89182
1.00000	3.09104	3.43399	3.33220	3.73767
1 00000	3.29584	3.21887	3.38777	3.98898
1.00000	2.89037	3.21887	3.40120	3.21887
1 00000	3.97029	3.17805	3.47197	2.70805
1.00000	3.98898	3.36729	3.38439	3.40120
1 00000	3.91202	3.13549	3.25809	3.25809
1.00000	3.78419	3.09104	3.51750	3.66356
1 00000	3.78419	3.29584	3.49651	3.71357
1.00000	4.18965	3.43399	3.37074	3.85015
1 00000	3.66356	3.36729	3.31054	2.19722
1.00000	3.82864	3.43399	3.44681	3.33220
1 00000	3.87120	2.94444	3.51154	3.63758
1.00000	3.63758	3.61092	3.31054	3.33220
1 00000	4.09434	3.09104	3.49651	3.46573
1.00000	3.78419	3.29584	3.31418	3.58352
2.00000	3.82864	3.76120	3.15273	4.15888
2.00000	3.55535	3.46573	3.21887	4.47734
2 00000	3.91202	3.58352	3.25037	3.71357
2.00000	3.71357	3.40120	3.09104	3.85015
2 00000	4.04305	3.73767	3.32504	3.76120
2.00000	4.04305	3.52636	3.33220	3.95124
2 00000	3.87120	3.58352	3.21887	3.21887
2.00000	4.09434	3.63758	3.25809	4.17439
2 00000	3.95124	3.49651	3.29584	3.73767
2.00000	3.89182	3.58352	3.25809	4.14313
2 00000	3.89182	3.78419	3.24259	3.93182
A	3.91202	3.46574	3.29584	4.44265
B	3.89182	3.13549	3.58352	4.12713
C	3.78419	3.68888	3.28840	4.00733
D	3.97029	3.68888	3.26957	3.58352

1. 0=ADENOMA
2. 0=BILATERAL HYPERPLASIA

Z1=AGE
Z2=POTASIUM
Z3=CARBON DIOXIDE
Z4=RENIN

N1=30
N2=11

APPENDIX (4.2.C): HAEMOPHILIA DATA

	Z1	Z2
1.	178.	84.
1.	164.	43.
1.	120.	34.
1.	204.	80.
1	140.	50.
1.	68.	73.
1.	100.	37.
1.	107.	37.
1.	125.	31.
1.	170.	34.
1	118.	120.
1.	173.	79.
1.	135.	70.
1.	135.	82.
1.	120.	66.
1.	200.	38.
1.	73.	40.
1.	110.	60.
1	135.	79.
1.	83.	67.
2.	85.	117.
2.	120.	93.
2.	125.	116.
2.	125.	99.
2	86.	71.
2.	70.	63.
2.	110.	121.
2.	70.	81.
2.	60.	64.
2.	50.	81.
2.	60.	57.
2.	90.	98.
2.	57.	55.
2.	70.	95.
2.	50.	57.
2.	63.	88.
2.	82.	91.
2.	60.	94.
2	80.	132.
2.	60.	55.
2	87.	74.
2.	55.	65.
2.	70.	120.
A.	162.	42.
B.	140.	83.
C.	105.	68.
D.	156.	146.
E.	120.	117.
F.	100.	127.
G.	90.	73.

Z1=FACTOR VIII RELATED ANTIGEN
Z2=FACTOR VIII ACTIVITY

1. =CARRIERS (N1=20)
2. =CONTROLS (N2=23)

Appendix (4.3): Testing Equality of Covariance Matrices

For $\underline{x}_{ij} \sim Np(\underline{\mu}_i, \Omega_i)$ $j=1,n_i$, we consider the hypothesis,

$$Ho: \Omega_1 = \Omega_2 \quad Vs \quad H_1: \Omega_1 \neq \Omega_2$$

where Ω_i, $i=1,2$ are the two population covariance Matrices from where we sample n_1, n_2 observations respectively.

When Ho is true:

$$\text{Let } S = \frac{1}{(n_1+n_2-2)} \sum_{i=1}^{2} S_i \text{ where } S_i = \sum_{r=1}^{n_i} (\underline{x}_r - \bar{\underline{x}})(\underline{x}_r - \bar{\underline{x}})^T$$

= sum of squares and cross-product matrices

then S is the estimate of the common covariance matrix.

$$\text{Let } M = (n_1 + n_2 - 2) \ln|S| - \sum_{i=1}^{2} (n_i-1) \ln \left|\frac{S_i}{n_i-1}\right|$$

$$C = 1 - \frac{2p^2 + 3p-1}{6(p+1)} \left[\sum_{i=1}^{2} \frac{1}{(n_i-1)} - \frac{1}{\sum_{i=1}^{2}(n_i-1)}\right]$$

Box (1949) showed that

$$MC \dot{\sim} \chi^2 \left[\frac{1}{2} p(p+1)\right] \text{ as the } n_i \text{ become large.}$$

The approximation is good if p does not exceed 4 or 5, and each n_i is perhaps 20 or more.

Appendix 5.1

Profile-likelihood, equal covariance case

We now briefly give the main 'equations' that we get when $\Omega_1 = \Omega_2$. The derivation uses the same methods of section (5.4, B).

Firstly, $\hat{\underline{\mu}}_1 = \dfrac{n_1\bar{\underline{x}}_1 + \lambda\underline{x}}{n_1 + \lambda}$, $\hat{\underline{\mu}}_2 = \dfrac{n_2\bar{\underline{x}}_2 - \lambda\underline{x}}{n_2 - \lambda}$

$$\hat{\Omega} = \frac{U}{n_1 + n_2}$$

where $U = \sum_{ij}(\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)^T + \dfrac{\lambda n_1}{(n_1 + \lambda)}(\underline{x}-\bar{\underline{x}}_1)(\underline{x}-\bar{\underline{x}}_1)^T$

$$- \frac{\lambda n_2}{(n_2-\lambda)}(\underline{x}-\bar{\underline{x}}_2)(\underline{x}-\bar{\underline{x}}_2)^T$$

The Lagrangian form is,

$L(\hat{\underline{\eta}}(\lambda), \lambda)$

$$= \frac{-(n_1+n_2)}{2}\log\left|\frac{U}{n_1+n_2}\right| - \frac{1}{2}\,tr\,[(n_1 + n_2)Ip]$$

$$- \lambda C + constant$$

Note: $|U|$ can be simplified by using,

$$|A + \underline{b}\,\underline{c}^T + \underline{d}\,\underline{e}^T|$$

$$= |A|\,[(1 + \underline{c}^TA^{-1}\underline{b})\,(1 + \underline{e}^TA^{-1}\underline{d}) - (\underline{e}^TA^{-1}\underline{b})(\underline{c}^TA^{-1}\underline{d})]$$

$$L(\hat{\underline{\eta}}(\lambda),\lambda) = \frac{(n_1+n_2)}{2}\,p\,\log(n_1+n_2) - \frac{(n_1+n_2)}{2}\log|S|$$

$$- \frac{(n_1+n_2)}{2}\log\left[1 + \frac{\lambda n_1}{(n_1+\lambda)}a_1 - \frac{\lambda n_2}{(n_2-\lambda)}a_2 - \frac{\lambda^2 n_1 n_2}{(n_1+\lambda)(n_2-\lambda)}\right.$$

$$\left.[a_1a_2 - a_{12}^2]\right]$$

$$-(n_1 + n_2)p/2 - \lambda c + constant$$

where,

$$S = \sum_{ij}(\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)^T$$

$$a_i = (\underline{x} - \bar{\underline{x}}_i)^TS^{-1}(\underline{x} - \bar{\underline{x}}_i) \quad (i=1,2)$$

$$a_{12} = (\underline{x} - \bar{\underline{x}}_1)^T S^{-1} (x - \bar{\underline{x}}_2)$$

$$\text{constant} = -(n_1+n_2)\, p \log(2\pi)/2$$

To show that $L(\hat{\underline{\eta}}(\lambda), \lambda)$ is convex it is sufficient to show that,

$$\left[1 + \frac{\lambda n_1}{(n_1+\lambda)} a_1 - \frac{\lambda n_2}{(n_2-\lambda)} a_2 - \frac{\lambda^2 n_1 n_2}{(n_1+\lambda)(n_2-\lambda)} [a_1 a_2 - a_{12}^2]\right] = T(\lambda) \text{ say}$$

is concave. We do so term by term. Note that a_1, a_2 and $[a_1 a_2 - a_{12}^2]$ are positive constants.

Thus $\frac{\delta^2}{\delta\lambda^2}\left[\frac{\lambda}{n_1+\lambda}\right] = \frac{-2n_1}{(n_1+\lambda)^3}$, clearly negative if $(n_1+\lambda) > 0$ $\Rightarrow$ concave (*)

$$\frac{\delta^2}{\delta\lambda^2}\left[\frac{-\lambda}{n_2-\lambda}\right] = \frac{-2n_2}{(n_2-\lambda)^3}, \text{ again concave if } n_2-\lambda > 0 \quad (**)$$

$$\frac{\delta^2}{\delta\lambda^2}\left[\frac{-\lambda^2}{(n_1+\lambda)(n_2-\lambda)}\right] = \frac{-2\, h(\lambda)}{(n_1+\lambda)^3 (n_2-\lambda)^3}$$

where $h(\lambda) = (n_2-n_1)\lambda^3 + 3n_1 n_2 \lambda^2 + n_1^2 n_2^2$.

It can easily be shown graphically that for all n_1, n_2 where $(-n_1 < \lambda < n_2)$, $h(\lambda) > 0$.

$$\Rightarrow \frac{-\lambda^2}{(n_1+\lambda)(n_2-\lambda)} \text{ is concave} \qquad (***)$$

Further, if f(x) and g(x) are both concave functions, then f(g(x)) is concave.

Using this result together with (*), (**) and (***) it is now clear that $L(\hat{\underline{\eta}}(\lambda), \lambda)$ is convex and we will have a unique minimum with respect to λ.

Finally, in a numerical algorithm to find λ_{max} of the profile likelihood we give the range of possible λ-values. These values can be obtained from the constraint $T(\lambda) > 0$ [Since $T(\lambda)$ is the argument of a logarithm]

$T(\lambda) > 0$ is equivalent to $g(\lambda) > 0$ where

$$g(\lambda) = (n_1+\lambda)(n_2-\lambda) + (n_2-\lambda)\lambda n_1 a_1 - (n_1+\lambda)\lambda n_2 a_2 - \lambda^2 n_1 n_2 [a_1 a_2 - a_{12}^2], \quad -n_1 < \lambda < n_2$$

Clearly $g(\lambda)$ is a quadratic in λ with a maximum.

We therefore need to solve $g(\lambda) = 0$ or solve,

$$-\lambda^2 r_0 + \lambda r_1 + r_2 = 0$$

where $r_0 = 1 + n_1a_1 + n_2a_2 + n_1n_2(a_1a_2 - a_{12}^2)$

$$r_1 = n_2 - n_1 + n_1n_2a_1 - n_1n_2a_2$$

$$r_2 = n_1n_2$$

The roots of this quadratic will give the range of possible λ-values.

Appendix (5.2)

The success of the Lagrangian method in solving the constrained optimisation problem can be illustrated with a geometric argument, following the idea of Whittle (1971). Firstly, define the (e,y) space where y represents the space into which $\underline{\eta}$ is mapped by the log-likelihood function. We need only consider the nature of the profile of the log-likelihood function, that is the upper boundary of the projection of the log-likelihood function into the (e,y) space, see figure (A).

The Lagrangian function is,

$$L(\underline{\eta}, \lambda) = \ell(e, \underline{\eta}) + \lambda(e-C).$$

In the (e,y) plane the equation of a line with slope λ can be written as

$$r = y + \lambda(e-C) \qquad (*)$$

In section (5.4.B) the solution to the Lagrangian problem involved maximising $L(\underline{\eta}, \lambda)$ with respect to $\underline{\eta}$. In the (e,y) plane this can be interpreted as finding the line of the form (*) for fixed λ which has maximum intercept with the line $e=C$.

If the upper boundary of the profile log-likelihood is concave, as shown in figure (A), then

$$L(\hat{\underline{\eta}}(\lambda), \lambda) = y + \lambda(e-C)$$

will be a "supporting" tangent for any given λ. In this case the minimisation procedure with respect to λ is equivalent to "rolling" the tangents for different λ's on the surface of the profile until the constraint $e=C$ is satisfied.

Figure (B) shows a case where the profile log-likelihood is not concave. The consequence is that there will exist a λ such that at least two values of $\hat{\underline{\eta}}$ (i.e. two sets of $\hat{\mu}_1, \hat{\mu}_2, \hat{\Omega}_1, \hat{\Omega}_2$)) maximise $L(\underline{\eta}, \lambda)$. Clearly here neither of these values will satisfy the constraint $e=C$, In this case therefore the Lagrangian approach will

not provide the correct solution to the constrained optimisation problem.

This would seem to suggest that multiple solutions to the maximisation of $L(\underline{\eta}, \lambda)$ with respect to $\underline{\eta}$ will indicate a non-concave profile. As our estimates (i.e. (5.4.2)) are unique we are fairly confident that this method will work in general. Again we note that we can always check empirically whether $\theta = C$ at our solution and hence whether we have the correct solution.

Obviously this argument requires some "tidying up" before it could be thought of as a formal proof.

Appendix (5.2)

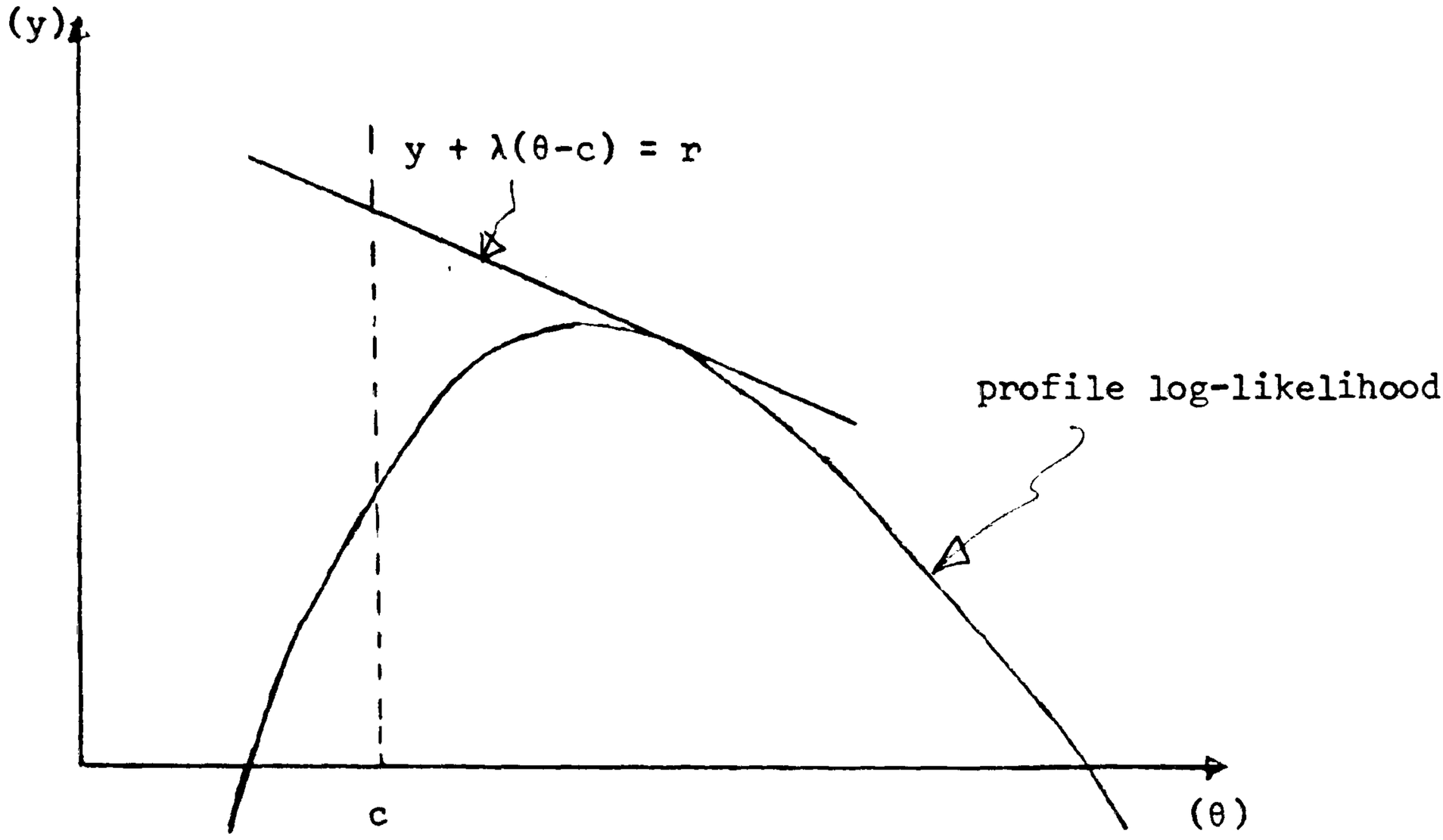


FIGURE (A): Concavity of the profile log likelihood, and the existence of "supporting hyperplanes"

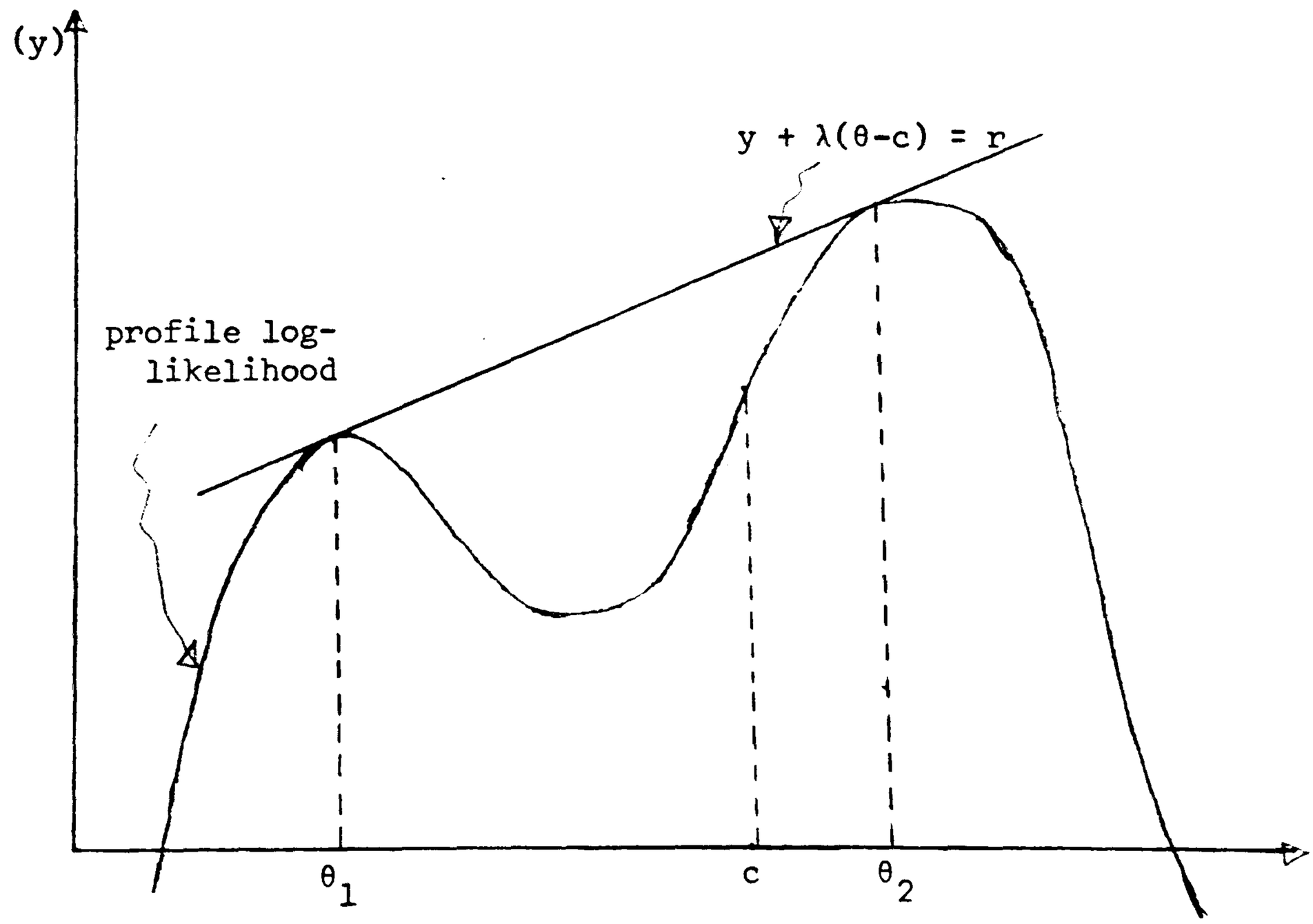


FIGURE (B)

REFERENCES

References marked with a (#) are not directly used but are useful for the general theory and methods of Discriminant Analysis.

Abramowitz, M and Stegun, I.A. (1972). Handbook of Mathematical Functions, New York, Dover Publications.

Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis, London, Cambridge University Press.

Aitchison, J. and Habbema, J.D.F. and Kay, J.W. (1977). "A critical comparison of two methods of Statistical Discrimination". Applied Statistics, 26, 15-25.

Ambergen, A.W. and Schaafsma, W. (1983). "Interval Estimates for Posterior Probabilities in a Multivariate Normal Classification Model" Printed at the Mathematical Centre, Kruislaan 413, Amsterdam.

Ambergen, A.W. and Schaafsma, W.(1984) "Interval Estimates for Posterior Probabilities, Applications to Border Cave". Multivariate Statistical Methods in Physical Anthropology, edited by G.N. van Vark and W.W. Howells (Reidel Publishing Co.).

Anderson, J.A. (1972). "Separate sample logistic discrimination". Biometrika, 59,1, 19.

Anderson, T.W. (1958). "An Introduction to Multivariate Statistical Analysis". Wiley, New York.

(#)Anderson, T.W. and Bahadur, R.R. (1962). "Classification into Two Multivariate Normal Distributions with Different Covariance Matrices". Ann.Math.Stat. 420-431.

Baker, R.J. and Nelder, J.A. (1978). Manual for the GLIM system, Release 3.

Bernardo, J.M. (1976). "Psi (Digamma) Function". Applied Statistics, 25, 315-317, Algorithm AS 103.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). "Discrete Multivariate Analysis: Theory and practice". MIT Press, London.

Box, G.E.P. (1949). "A General Distribution Theory for a class of likelihood criteria". Biometrika, 36, 317-346.

Chatterji, S.D. (1963). "Some elementary characterizations of the Poisson distribution". American Math.Monthly, 70, 958-964.

Cox, D.R. (1970). "Analysis of Binary Data". Methuen's Statistical Monographs, London.

Critchley, F. and Ford, I. (1984). "On the covariance of two non-central F random variables and the variance of the estimated linear discriminant function". Biometrika 71,3 637-638.

Critchley, F. and Ford, I. (1985). "Interval estimation in discrimination: the multivariate normal equal covariance case". To be published in Biometrika 1985, 72,1.

Dixon, W.J. (1977). B.M.D.P. Biomedical Computer programs. University of California press.

Eadie, A.S., Horton, P.W., Adams, F.G. and Hilditch, T.E. (1980). "Comparison of Nuclear Imaging Formats using R.O.C. Analysis. Unpublished paper. (Dept. of Nuclear Medicine, Western Infirmary, Glasgow).

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. Annals of Statistics, 7,1 1-26.

Efron, B. (1981). Non-parametric standard errors and confidence intervals. The Canadian journal of Statistics, 9,2 139-172.

Efron, B. (1982). The Jackknife, the bootstrap and other resampling plans. SIAM monograph #38, CBMS-NSF.

Efron, B. and Gong, G. (1983). A leisurely look at the Bootstrap the Jackknife and Cross-Validation. The American Statistician, 37,1.

Everitt, B.S. (1977). "The Analysis of Contingency Tables" Monographs on Applied Probability and Statistics. Chapman and Hall.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Ann.Eugen. 7, 179-188.

(#)Gilbert, E.S. (1969). "The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. Biometrics, 25, 505-516.

Graybill, F.A. (1983). "Matrices with applications in Statistics". Wadsworth.

Gupta, S.D. (1968). "Some Aspects of Discrimination Function Coefficients". SANKHYA (A), 30, 387-400.

Guttman, I. (1970). "Statistical Tolerance Regions: Classical and Bayesian". No.26, Griffin's Statistical Monographs and Courses.

Houston, A.S. and Macleod, M.A. (1979). "An intercomparison of computer assisted image processing and display methods in liver Scintigraphy. Physics in Medicine and Biology, 24, 559-570.

Johnson, N.L. and Kotz, S. (1970a). Distributions in Statistics. "Continuous univariate distributions, Vol.1". Houghton Mifflin Company, Boston.

Johnson, N.L. and Kotz, S. (1970b). Distributions in Statistics. "Continuous univariate distributions, Vol.2". Houghton Mifflin Company, Boston.

Johnson, N.L. and Kotz, S. (1970c). Distributions in Statistics. "Continuous Mutlivariate distributions". Houghton Mifflin Company, Boston.

Kalbfleisch, J.G. (1979). "Probability and Statistical Inference II". Springer-Verlag, New York.

Kalbfleisch, J.G. and Sprott, D.A. (1970). "Application of Likelihood Methods to Models involving Large Numbers of Parameters". Journal of the Statistical Society, B32, 175.

Kendall, M.G. and Stuart, A. (1966). The Advanced Theory of Statistics, 3, London.

(#)Krzanowski, W.J. (1977). "The Performance of Fisher's Linear Discriminant Function under Non-Optimal Conditions". Technometrics 19,2 191-199.

Lachenbruch, P.A. (1975). "Discriminant Analysis". Hafner Press.

(#)Lachenbruch, P.A., Sneeringer, C. and Revo, L.T. (1973). "Robustness of the Linear and Quadratic Discriminant Function to Certain Types of Non-Normality". Communications in Statistics 1,1, 39-55.

Lusted, L.B. (1971). "Decision making studies in patient management". New England Journal of Medicine, 284, 416-424.

Lusted, L.B. (1978). "General problems in medical decision making with comments on R.O.C. analysis". Seminars in Nuclear Medicine, 8, 299-306.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). Multivariate Analysis, Academic Press, London.

(#)Marks, S. and Dunn, O.J. (1974). "Discriminant Functions when Covariance Matrices are Unequal". Journal of the American Statistical Association, 69, 555-559.

McCullagh, P. and Nelder, J.A. (1983). "Generalized linear Models". Monographs on Statistics and Applied Probability, Chapman and Hall.

Metz, C.E. (1978). "Basic Principals of R.O.C. Analysis". Seminars in Nuclear Medicine, 8, 283-298.

Moran, M.A. and Murphy, B.J. (1979). "A closer look at two alternative methods of Statistical Discrimination". Applied Statistics, 28,3, 223-232.

NAG (1978). Manual for the NAG package, Mark 7. Numerical Algorithms Group, Oxford.

Nelder, J.A. and Wedderburn, R.W.M. (1972). "Generalised Linear Models". J.R.Statist.Soc.A., 370.

Prentice, C.R.M., Forbes, C.D., Morrice, S. and McLaren, A.D. (1975). "Calculation of Predictive Odds for Possible Carriers of Haemophilia". Thrombos.Diathes.Haemorrh. (Stuttg), 34, 740.

(#)Remme, J., Habbema, J.D.F. and Hermans, J. (1980). "A Simulative Comparison of Linear, Quadratic and Kernal Discrimination". J.Statist.Comput.Simul. 11, 87-106.

Rigby, R.A. (1982). "A credibility interval for the probability that a new observation belongs to one of two multivariate normal populations". J.R.Statist.Soc.B. 44, 212-220.

Schaafsma, W. and Van Vark, G.N. (1977). "Classification and Discrimination problems with applications, Part I, Statistica Neerlandica, 31.

Schaafsma, W. and Van Vark, G.N. (1979). "Classification and Discrimination problems with applications, Part IIa, Statistica Neerlandica, 33.

Siskind, V. (1972). "Second moments of inverse Wishart-matrix elements". Biometrika, 59,3 690-691.

Snedecor, G.W. and Cochran, W.G. (1967). "Statistical Methods". Sixth Edition. Iowa State University Press.

(#)Wahl, P.W. and Kronmal, R.A. (1977). "Discriminant Functions when Covariances are Unequal and Sample Sizes are Moderate". Biometrics 33, 479-484.

Whittle, P. (1971). Optimization under Constraints. Wiley Publications.