A STUDY OF TWO PRACTICAL PROBLEMS

IN CLINICAL TRIALS METHODOLOGY

by

JANET G. FINDLAY

A Dissertation submitted to the

UNIVERSITY OF GLASGOW

for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics

April 1990.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

ACKNOWLEDGEMENTS

First, I wish to thank my supervisor, Dr Gordon D. Murray, for his guidance and encouragement throughout the period of my research, and also for providing such a friendly working environment.

I must also thank my friends, especially those in the University Departments of Statistics and Surgery, for the support they have so freely given over the last five years. In particular, I would like to express my gratitude to Dr J.W. Kay of the former for consistently looking after my spiritual well-being.

I wish also to thank Janssen Pharmaceutica, Beerse, Belgium for sponsoring this research.

Finally, I would like to express my heartfelt gratitude to my family, without whose support this research would have been made infinitely more difficult.

DECLARATION

Some of the one-sample results appearing in Section 4.1 have appeared previously in Murray and Findlay(1988).

TABLE OF CONTENTS

FIGURES

TABLES

SUMMARY

In this thesis, two problems which commonly arise in the context of clinical trials, but which are often dealt with inappropriately are considered. The thesis consists of two parts, one corresponding to each of the problems considered.

Part I (Chapters 2-4 inclusive) considers the problem of how to handle incomplete multivariate data, while Part II (Chapters 5-7 inclusive) looks at some methods for comparing groups which have an a priori ordering.

Chapter 1 provides a general introduction to clinical trials methodology, describing some of the basic concepts involved, for example randomisation and blinding. In addition, some recent developments, for instance meta-analysis, are outlined. The chapter ends with an informal review of the last two years' issues of two leading medical statistics journals, namely "Statistics in Medicine" and "Controlled Clinical Trials", in order to highlight some of the topics of current interest.

Chapter 2 introduces the problem of handling incomplete data. The basic terminology for describing missing data mechanisms is given, with special emphasis being given to the ideas of "Missing at Random" as described in Rubin(1976). The chapter ends with a review of some of the commoner methods of analysis used. Some of the potential problems which can arise when using these methods are outlined.

In Chapter 3, the ideas of maximum likelihood estimation in the presence of incomplete data are developed. After describing some existing likelihood-maximisation techniques for handling single groups in isolation, these ideas are then extended to the multiple-group problem. Methods are developed for fitting models where several groups are constrained to have a common covariance matrix, but are allowed to have different means. This allows emulation of the likelihood ratio testing procedures usually only applicable if the data are complete.

In Chapter 4, the techniques of the previous two chapters are applied to three real examples. Some of the work of Section 4.1 has appeared previously in Murray and Findlay(1988).

The main results of chapters 2-4 are summarised. Emphasised is the fact that inappropriate handling of even a relatively small

amount of incomplete data can have a substantial effect on the results obtained.

Chapter 5 marks the beginning of Part II of the thesis. The ideas of order-restricted inference are introduced, and some potential areas of application are outlined. After a review of some of the existing literature, it becomes clear that there exist certain common clinical trials scenarios where no appropriate testing procedures are available. Notably, there is no procedure available for testing for differences between ordered groups of Normally distributed data while incorporating covariate information. After devising a suitable testing technique for such problems, eight different tests are defined which incorporate differing degrees of information and assumptions about the data under study, in terms of distributional assumptions, covariate information and prior group ordering information.

In Chapter 6 the aim is to compare the performance of the eight tests defined in Chapter 5 under a variety of conditions, after evaluating and optimising the specifications of one relatively recent test (Marcus and Genizi(1987)).

The performance of the eight tests are assessed :

(a) as the group separation is varied,

(b) as the covariate/response correlation is varied, and

(c) as the error distribution is made other than Normal.

The conclusions drawn are that while the incorporation of covariates and ordering information is always worthwhile, the assumption of Normality does not greatly affect the sensitivity of the analyses performed. It is noted that there could be some difficulties in gaining acceptance by clinicians for the more complicated procedures described, and that, in practical terms, gaining statistical sensitivity could well be at the cost of losing credibility.

In Chapter 7, ordered testing procedures are applied to some published data sets, and the results thus obtained are compared to those which were published.

Finally, Chapter 8 outlines the conclusions for the thesis as a whole, stressing the important implications as regards clinical research. Some ideas for future research in the topics covered are proposed.

Chapter 1 - Introduction

In the dynamic pharmaceutical industry, new products are continuously being developed, for which there are often made claims of improved efficacy over existing products. Behind such statements lie years of research into the products' actions, including side-effects, involving Clinical Trials.

Several good general texts are available describing the philosophy behind clinical trials and also their execution, for example Pocock(1983), Meniert(1986), Shapiro and Louis(1983), Schwartz et al(1980).

1.1 : What is a Clinical Trial ?

A clinical trial is a planned experiment designed to study, in man, the effects of two or more medical interventions, where one of these interventions might well be some new "test" treatment. Of interest tend to be questions like :

"Does the proposed treatment produce the desired effect ?"

or "What would be the most appropriate treatment for future patients with a given medical condition ?"

The American Food and Drug Administration (F.D.A.) has classified clinical trials for drug development into four categories or "phases", as defined below, these phases all occurring after initial animal experimentation into the general drug safety.

Phase I : Such experiments represent the first exposure in man to the treatment under study. Here, the aims are to establish, principally, the drug safety and tolerability (in terms of side-effects), rather than its efficacy. Information is gained on the drug metabolism, pharmacokinetics, duration of action, etc.

Usually, Phase I trials are conducted in normal, healthy volunteers, rather than in those types of individual to whom the treatment would, ultimately, be applied.

Phase II : Such experiments tend to be relatively small, and aim to study drug efficacy and safety. At this stage, a new

treatment, possibly at a variety of doses, is generally compared to a placebo (dummy) treatment, and assessment is made of the new treatment's potential.

Phase III : Unlike Phase II, comparison is generally made against the standard treatment(s) used, with a greater number and wider range of patients, under conditions simulating the intended conditions of use, in order to see if the treatment's potential is realised in clinical practice.

Phase IV : This final phase takes the form of post-marketing surveillance, monitoring the longer-term implications of the treatment including its efficacy, toxicity, its effects on mortality, and, in particular, rare side effects which would not be picked up in the Phase II and III trials.

Note that the distinctions between the Phases are not as clear-cut as the above paragraph would suggest. Although the general chronological order of the main features is as described, it might be difficult to distinguish between a "late Phase II" and an "early Phase III" trial.

Throughout the following work, only Phase II and III trials will be given any emphasis.

Before commencing any clinical trial, clearly the most important question is, "Which treatment has to be assessed ?". Having established the treatment of interest, this gives an indication as to which class of patients are going to be eligible for inclusion into the study. For example, if the aim was to assess a new anti-hypertensive agent, clearly the patients of interest would be those who had some form of high blood pressure.

Having a clear definition of the aims of the study would also identify a broad class of responses which could be of interest. In the situation above, an indicator of blood pressure reduction would probably be an appropriate response measure. This could be simply the magnitude of blood pressure reduction, or, alternatively, criteria could be set out in advance defining blood pressure "control" - e.g. an individual could be defined to have "controlled" blood pressure if their systolic blood pressure was lower than 140 mmHg and their diastolic blood pressure was lower than 95 mmHg (and "uncontrolled" otherwise).

1.2 : Treatment Assessment

One obvious point which should hardly need to be stated is that all patients are different : otherwise there would be little need for clinical trials. The same medical condition in two individuals will not necessarily follow the same course, and through time, even the condition of a single patient may either deteriorate or stay about the same or even spontaneously improve to the point of "cure".

As a result, any improvement seen in a person's disease state during the course of a clinical trial may not necessarily be due to the applied treatment.

This is the reason why, in clinical trials, it is necessary to obtain results under control conditions, to act as a reference point for results obtained under test conditions. Note that the control treatment may take the form of either a standard treatment or a placebo (dummy) treatment. Such control results can originate from many different sources.

In Parallel Group Studies, the control and test results are evaluated concurrently in separate, independent, groups of individuals.

In Crossover Studies, each individual receives sequentially the control and test treatments under similar conditions. A random mechanism (see later) is used to dictate the order in which an individual receives the control and test treatments. Careful study design ensures a balance of the number of cases receiving the test treatment first and the number of cases receiving the control treatment first.

In Self-Controlled Studies, each case acts as their own control. A single treatment is assessed by comparing the results obtained under that treatment to those obtained in the same cases at some point (not specified) when no treatment was being applied.

In Studies with Historical Controls, control results are obtained from previous cases who happened to have been given a standard treatment. This can be problematic if there have been changes in the patient population through time, or if the type of cases who were given the standard treatment were systematically

different in some way. For example, certain treatments are not recommended for use in the elderly. If the standard treatment had such restrictions, but the new treatment did not, this could lead to comparisons of the treatments being biased due to these differing age distributions.

1.3 : Treatment Assignment

1.3.1 : Treatment Assignment in Parallel Group Studies

Clearly, for parallel group studies, there have to be means by which individuals are assigned to the different groups. One simple way would be either to allow people to select their own treatment, or to allow the clinician involved to perform the selection.

However, this can, again, lead to systematic differences in the cases in the different groups. For example, the investigator might favour one treatment for the more severe cases (probably the non-control treatment, or the treatment the perceived to be "better") and another for the less severe cases. This could lead to the new treatment being shown in a misleadingly unfavourable light, since the most severely ill cases would, generally, be expected to fare less well than their healthier counterparts.

A family of methods of case assignment which could be thought of as an improvement on the judgemental methods of assignment above, would be the family of Systematic Methods of Case Assignment. Some examples of systematic assignment are described below. Throughout, let A and B represent assignment to treatments A and B, respectively.

Alternate Assignment : Cases are alternately assigned to the treatment groups, so that looking at the assignments of the first few cases, we would have A B A B A B A B ... .

Assignment by Date of Birth : If day of birth EVEN , assign to treatment A, and otherwise assign to treatment B.

Alphabetic Assignment : Assign using the initial letter of the first name : this has been shown to introduce biases, since certain ethnic groups are over-represented for certain letters of the alphabet.

One real problem with each of these systematic methods is that

the assignment for any particular patient is predictable. This could lead to biases, again, if the investigator had a preference for one or other treatment for a given patient, since if the patient was not going to receive the preferred treatment, then the investigator could choose not to enter the patient into the study at that particular time.

It begins to emerge that it would be better if patients could be assigned to their groups in some objective, unpredictable way, i.e. it would be better to use some form of randomisation (random assignment) procedure. In its simplest form, where there are two groups and approximately 50% of the cases are to be assigned to each group, assignment could be based on the result of tossing a coin (e.g. If a Head, assign to A and if a Tail, assign to B). For more complicated situations, assignment procedures could be devised based on using tables of random numbers. The correct use of randomisation means that there is no bias in the patient selection for the different treatment groups, and so the results obtained can be more reliably assessed.

1.3.2 : Methods of Randomisation for Parallel Group Studies

Simple randomisation procedures as described in the previous paragraph have certain advantages and disadvantages compared to more complicated schemes. Firstly, by virtue of their simplicity, errors are less likely to occur in their application than they are for more complicated procedures. Also, in the long term, approximately the correct proportions of individuals will be assigned to the different groups. However, in the short term, there can sometimes be produced fairly serious imbalances in the numbers of cases assigned to the different groups.

To surmount this problem, two alternative approaches have been devised : the Random Permuted Blocks design and "Biased Coin" approaches.

In the Random Permuted Blocks Design, if we let K represent the number of treatments to be assessed, patients are assigned in blocks of mK. Within each block, each treatment is assigned a total of m times. Thus, within each block, the treatment assignments are exactly balanced. A value of m should ideally be chosen to be large enough such that the next assignment is difficult to predict.

In "Biased Coin" Approaches (see Efron(1971)), before each patient is randomised, the current balance of treatment allocations is considered. Whichever treatment has had fewest allocations so far is assigned to the next case with a probability of greater than 0.5.

At times, however, merely balancing groups with respect to sample size will not be enough. Often, for each patient there will be certain prognostic factors known to affect the outcome (age, for example), and these should be balanced, as far as possible, in the different treatment groups. Simple randomisation procedures do not guarantee such balance. Here it is advisable to use techniques involving stratification, such as those described below.

Random Permuted Blocks Within Strata : Patients are divided into strata according to the prognostic factors known to be relevant. For example, patient ages might be divided into Young, Middle-Aged and Old strata. Within each stratum, the treatment allocations are balanced as in the case of the Random Permuted Blocks design. This allocation method is useful as long as either there are only few strata or the study is very large.

Minimisation : This method is useful for matching groups for a large number of prognostic variables. Each patient allocation is made by a random mechanism , but biased so as to make the groups as well matched as possible. For example, if, at a particular stage in a study, patients assigned to A were, on average, older than patients assigned to B, then a young recruit would be assigned to A with a probability greater than 0.5.

1.3.3 : Randomisation in Crossover Studies

Recall that in crossover studies, each case receives more than one of the treatments under consideration, so that within-patient comparisons are possible, rather than just between-patient comparisons as in the case of parallel group studies.

Such studies, where appropriate, can increase the precision of treatment comparisons, by eliminating the factor of inter-individual variability from the treatment comparisons.

It must be noted, however, that crossover studies are not always possible. For example, if one wants to compare the

long-term effects of treatments, it will not be possible to assess such effects for several treatments within the same individual. Crossover studies are really only appropriate where short-term responses to therapy are being studied in patients with a stable chronic disease state, for example in essential hypertension.

Time plays an important role in crossover studies :

How long should the period on each treatment be ? Ideally, one would want to allow sufficient time for each treatment to reach its maximal effect in the course of their period of study.

How long should be allowed between the first and second treatment periods ? A sufficiently long washout period should be allowed so as to let the effect of the treatment in the first period disappear, or at least reduce to a minimal level, before the second period commences.

What if there is a time effect on the disease state ? It could occur that patients would tend to be admitted to a trial when their disease state was particularly bad, as well as there probably existing an underlying progressive disease state. It could easily occur, then, that the condition of patients in Period 1 would be systematically different from their condition in Period 2. Such an effect would be confounded with the treatment effects unless some cases received treatment A in period 1 and some received treatment B in period 1 (confining consideration only to the simple two treatment, two period case).

As in the parallel group situation, the assignment of cases to their treatment regimen should be made using randomisation. Here there are two regimens to be considered : A followed by B (A-B) and B followed by A (B-A).

When using such study designs, as well as estimating the differences between the treatment effects, it would also be desirable to be able to estimate such factors as period effects, group effects and treatment by period interactions. However, in the simple two treatment, two period study, certain groups of parameters are confounded with each other, i.e. they may not be estimated separately. For example, the treatment by period interaction and the difference between the carry-over effects cannot be estimated separately. Some of the problems of confounding can be avoided either by carrying out the treatment comparisons over three periods instead of two (e.g. by repeating

the second period treatment : see Morrey(1985)) or by taking run-in and wash-out readings into account (see Jones and Kenward(1989)).

It should be noted that the use of crossover designs can sometimes eliminate the need for stratification - the cases receiving the different treatments are often automatically balanced with respect to most prognostic factors likely to be of interest.

No specific details of the analysis of crossover studies will be given here, since the problem has been well-documented elsewhere, in, for example, Hills and Armitage(1979), Grizzle(1965), Kershner and Federer(1981), Jones and Kenward(1989).

It would be misleading to think of randomisation as some kind of "cure-all" for the problems of bias in clinical trials. Even using the best of randomisation methods, it is still easy to obtain distorted results if the administered treatment is identifiable by either the investigator or the patient.

For example, if the patient knows he is receiving a standard treatment rather than the new "revolutionary" treatment, psychological factors could adversely affect his results.

If the investigator knows that a patient is receiving the new, perceived to be better, treatment, he may convey this, possibly indirectly, to the patient. Again, this could have some effect on the results.

Such problems can sometimes be avoided using blinding.

1.4 : Blinding

In a single-blind procedure, the patient is not informed of which treatment is being given. In a double-blind procedure, neither the investigator carrying out the patient assessment nor the patient knows the treatment assignment. However, such blinding is not always possible.

If the aim was to compare two forms of medical intervention, one involving drug therapy, and the other involving surgery, clearly it would not be possible to "blind" the patient to the effects of surgery! Neither would it be possible, ethically, to

subject patients to "sham" surgery, just so as to mimic the visible effects of the alternative treatment - the mechanisms of blinding should certainly never put patients at risk unnecessarily.

In order that blinding be effective, there must be no obvious differences between the competitive treatments. However, what can be done if there is no effective rival treatment, and comparison is to be made to a "no treatment" alternative? This is where the use of placebo control is required.

What is a Placebo ? A placebo is a totally inert treatment given as a substitute for an active treatment. Desirable properties of a placebo would be that it should :

(i) look like the active treatment (size, shape, colour, etc.)
(ii) taste like the active treatment
(iii) smell like the active treatment
(iv) be administered by the same route and with the same frequency as the active treatment.

Although the constituents of a placebo are inert, patients will still usually react in some way, purely due to the psychology of being treated. Often, dramatic improvements in the patient state can be observed. Also, it is common for side-effects to be reported.

It is natural to assume that such spontaneous changes in the patient's condition will also underly the observed responses to active treatment. It would thus give a truer indication of the "real" treatment effect if comparisons were made to a placebo, rather than to a no treatment alternative.

1.5 : Ethical Considerations

In 1960 the World Medical Association issued a set of guidelines to the ethical requirements of clinical research : The Declaration of Helsinki (revised in 1975).

Basically, it is unethical to conduct poorly planned research. More specifically, investigators should try to avoid :

(a) Bias : which could lead to exaggeration of the treatment effect.

(b) Using too few patients : if too few patients are used, the results obtained will be neither precise nor reliable.

(c) Not publishing their findings : if the aim of clinical research is, ultimately, to further medical knowledge, then clearly there is an ethical responsibility for investigators to publish their findings.

One important issue covered by the Declaration of Helsinki is that of informed consent. The Declaration states that in clinical research, the investigator should obtain the patient's written consent, where this is possible. Such consent must be obtained without duress, with the patient fully understanding the personal implications of the research. This means that they must be made aware of the possible side-effects and risks of treatment, as well as the possible benefits. They must also be informed if there is a possibility that they could receive an inert (placebo) treatment.

Some of the practical difficulties of obtaining a patient's informed consent to participate in a randomised study can be avoided by using the Randomised Consent Design (see Zelen(1979)). Instead of randomising patients to a given treatment group, they are randomised to one of the two groups below :

(i) Seek consent group

(ii) Do not seek consent group .

All patients in group (ii) receive the standard treatment, A, which they would usually receive, thus their consent is not required. Those patients in group (i) are offered the new treatment, B. If they accept, then they are given treatment B. Otherwise, they receive treatment A as usual. An important point about this design is that in the analyses, all of the cases in group (i) are compared to the cases in group (ii), rather than comparing purely cases on A to cases on B.

What results, rather than a comparison of treatment A with treatment B, is a comparison of the policy of offering treatment B to the policy of not offering treatment B.

Recently, much interest has been generated by the ideas meta-analysis. Here, sample sizes are artificially increased by pooling together the data from several similar small (individually uninformative) studies to produce a concensus view.

The reasoning behind meta-analysis is described in the following section.

Meta-Analysis (see e.g. Yusuf et al(1985), Murray(1990), Boissel et al(1989)) : Certain circumstances dictate the need for very large clinical trials. For example, to study conditions with low occurrence-rates, or in order to demonstrate a modest, but clinically relevant treatment difference, requires a surprisingly large number of patients. However, it is more common for a large number of small studies to be performed, often resulting in contradictory results, rather than one larger, adequate study.

Meta-analysis aims to draw reliable conclusions about a treatment benefit by integrating the findings of several similar small studies involving that treatment. Criticisms aimed at meta-analysis tend to be that even if trials are broadly similar in their aims, it is almost inevitable that there will be some differences in, for example, the patient populations under study, which could lead to systematically different results. What is hoped is that, generally speaking, the results will tend to be inclined in the same direction, even if the actual magnitude of the treatment effects are different. In order to make the pooling of trial results relatively straightforward, it is usually most appropriate to opt for uncomplicated measures of outcome, for example success/failure or alive/dead, so that essentially only the appropriate "event rates" need to be combined.

One of the most difficult aspects of meta-analysis is the choice of the small trials for inclusion. Ideally, all relevant trials, including aborted trials, would be included, but it is difficult to identify all of these trials, partly due to the lack of any formalised registration of research in progress (except in a very few areas such as cancer research). Reliance simply on published trial findings as a source of results is not sufficient, due to the well-known phenomenon of publication bias, whereby a trial leading to "positive" findings is more likely to be published than one leading to "negative" findings. Clearly, if a meta-analysis was performed only on the basis of published trials, a very distorted view of the treatment benefit could be obtained.

Also, bias can be introduced into a meta-analysis if the

results of non-randomised studies are included. The general concensus of opinion seems to be that it is best to restrict attention only to randomised studies in the area of interest, taking care to assess other possible bias-sources within each trial used e.g. All patients should be included in an Intention to Treat type of analysis, with all withdrawn patients accounted for and incorporated, as far as possible.

Various algorithms have been proposed for the assessment of trial quality e.g. the scoring system of Chalmers et al(1981). Chalmers et al argued that sensitivity analyses should be performed in order to assess whether trials of poorer quality could be distroting the meta-analysis results.

Despite the many difficulties which can arise in performing meta-analyses, they can sometimes provide valuable insight into problems where individual small trials are uninformative and very large trials are impractical.

An idea which has become popular relatively recently is that of the competition between individual and collective ethics. Ideally, each patient would receive the treatment of benefit to him, but this could be contrary to the aims of benefit to future patients - in order to make useful progression in medical knowledge, it could well be more efficient for the patient to receive the alternative treatment.

Another idea which has become popular recently is the distinction between two types of trial : explanatory trials and pragmatic trials (see Schwartz et al(1980)). These have quite different aims, and as a result are quite different in other aspects, including their limitations and also the possible conclusions which can be drawn.

Generally speaking, in drug development, early (Phase II) trials tend to be of the explanatory type, whereas later (Phase III) trials tend to be of the pragmatic type.

1.6 : A Comparison of Explanatory and Pragmatic Trials

1.6.1 : Explanatory Trials

The Aims : These aim to provide useful information at a biological level, perhaps looking at the mechanisms of a drug's action. The criteria for outcome assessment are chosen to be of biological importance, e.g. tumour regression.

Properties

(a) Useful for assessing whether a new drug has any effect at all, by comparison with no therapy.
(b) Patients are chosen to be as homogeneous as possible, and as likely as possible to respond to treatment.
(c) The aim is to maximise the biological effect.
(d) Treatments are compared using standard statistical tests.
(e) Experimental conditions are equalised as far as possible (e.g. the same method of administration is used for the different treatments, and these treatments are given in equal doses), using refined laboratory conditions.
(f) Patients violating the study protocol are omitted from analyses.
(g) The type of measurements used are objective, and are chosen to have biological importance e.g. biochemical measurements.

Type of Conclusions Possible : Valuable biological information may be obtained about drugs' actions, leading to advancement of scientific knowledge.

Drawbacks : Since the study has used a carefully chosen, homogeneous, responsive set of patients, the findings obtained may not be applicable to the intended population of future application. A positive result in the selected group will not necessarily imply positive results in general.

1.6.2 : Pragmatic Trials

The Aims : These trials compare treatments under conditions reflecting current clinical practice. Measures of outcome are chosen to be of direct relevance to the patient, for example

survival time taken in conjunction with quality of life during that time.

Properties

(a) Such trials are useful for assessing whether a new treatment is better than the treatment currently in use.

(b) Patients are chosen to be representative of those to whom the treatment would be applied in practice, rather than being chosen to be as homogeneous as possible.

(c) The aim is to optimise the patient benefit, i.e. to get a reasonable response without excessive side-effects.

(d) Analyses are based on decision theory, trying simultaneously to consider several different factors in order to select the best treatment overall.

NOTE : Although this is the natural consequence of the pragmatic approach as described by Schwartz et al(1980), such analyses are rarely used in practice.

(e) Conditions of administration might be chosen so that for each patient the drug is at its optimal dose level, rather than equalised doses being used as in the explanatory situation. These experiments are performed under as natural conditions as possible, rather than under restrictive laboratory conditions.

(f) Patient withdrawals can be incorporated, since one is only comparing the policies "A if possible" and "B if possible", rather than the purer hypotheses of "Treatment A" versus "Treatment B".

(g) Instead of simply looking at criteria with clear biologiocal interpretation, here one wants to provide an overall assessment of the competing treatments, taking into account such subjective factors as patient well-being as well as more directly measurable factors such as survival time.

Type of Conclusions Possible : Such studies do provide an answer to the general, practical, problem of interest, without restricting attention to one very specialised, probably atypical, subgroup of patients.

Drawbacks : Although a negative result in a large, well conducted, pragmatic study is useful as an indication that there

is no clinically relevant treatment difference overall, this does not exclude the possibility that, within certain patient subgroups there could be relevant differences. While such studies are useful for producing treatment recommendations, they do not generate leads for further research.

1.6.3 : The Choice Between Explanatory and Pragmatic Trials

From the above section, it appears that before performing a clinical trial, one would have to decide whether the aim was to solve some practical problem (leading to a pragmatic trial) or, alternatively, to advance scientific knowledge (leading to an explanatory trial). Certainly, some trials could be purely explanatory and others could be purely pragmatic. Examples of the former tend to occur in early Phase II trials, where the aim is purely to assess whether a new drug has any effect at all. Examples of the latter tend to occur in late Phase III trials, where the aim is purely to decide on the best patient management strategy.

More commonly, what is desired is some balance between these two extremes. Ultimately, the aim will be to further medical knowledge, but this cannot ethically be done in a way contrary to the good of the individual patient.

1.7 : Statistical Aspects of Clinical Trials

The roles of the statistician in the planning and execution of clinical trials are wide and varied. In fact, the involvement of the statistician should begin at the outset of the trial, and continue until its final termination and evaluation. To illustrate this, some of the tasks of the clinical trials statistician are outlined below.

Trial Design : Given the aims and objectives of a given study, it is for the statistician to decide which trial design is most likely to answer the questions of interest in the most efficient manner. Indeed, it is often the statistician who forces the clinicians to think critically about the exact aims and objectives of the trial.

Sample Size Calculations : Before the appropriate sample size can be calculated, certain information must be made available to the statistician :

(a) The main aim of the study, e.g. estimating a proportion or comparing two means or comparing two death rates.
(b) The chosen significance level.
(c) The minimum difference between the groups under study which is important to detect and
(d) With how much certainty (Power).
(e) An estimate of the underlying variability in the responses of interest.

Given these constraints, there are available standard statistical approaches for calculating the required sample size for the problem under consideration. Note that for the case of survival analyses, the number of expected deaths in the groups form the basis of the sample size calculations. Studies of conditions with a low death rate automatically require a greater number of cases to detect the same magnitude of treatment effect as in a similar study with a higher underlying death rate.

No specific details of sample size calculation will be given here, since much work has been done in this area, and the appropriate methodology for a wide range of situations has been widely documented (e.g. Machin and Campbell(1987)).

It should be noted that often, there will be financial constraints or time limitations on a study, so that there will be a ceiling on the number of patients which it is possible to study. In such cases, sample size calculations are commonly used in reverse, working from the maximum number of patients available to obtain an estimate of the resultant power to detect a given treatment difference - if the power is too low, then there is little point in performing the study at all, unless more limited objectives can be accepted, perhaps tipping the balance towards explanatory rather than pragmatic questions.

As mentioned earlier, much interest has been generated recently by the ideas of meta-analysis, where sample sizes are artificially increased by pooling together the data from several similar small studies to produce a concensus view.

Randomisation : Assuming that a trial has reached the point where the appropriate sample sizes are attainable, it is the role of the statistician to devise an execute an appropriate randomisation procedure.

Data Checking : From this point, the statistician takes on more of an administrative and regulatory role. As the data accumulates, careful checking is required, looking for any obvious administrative errors. For example, was a patient actually eligible for inclusion in the study at all? If so, are all of their data forms complete, and if incomplete, can any of the "blanks" be filled? Are there any obvious errors, e.g. an age of 156, or a categorical variable coded as "4" when the only valid codings are "0" and "1"? Are there any inconsistencies in the data, for example a date of birth occurring after a date of randomisation?

Monitoring the Trial Progress : One important task of the statistician is to monitor the study as regards drug safety and possible adverse drug reactions. Unexpected deaths or side-effects on any treatment must be investigated, especially if one group appears to have many more problems compared to the other groups under study. Depending on the circumstances, appropriate action in such cases can be as simple as withdrawing a particular treatment from certain "at risk" groups, or as radical as stopping the study altogether.

At the opposite end of the spectrum of trial monitoring, the statistician must, ethically, be able to stop a study if sufficient evidence has emerged to show beyond reasonable doubt that a given treatment is better than the alternatives - from that stage, it would be unethical to continue to randomise patients to the inferior treatment. To decide what would constitute "beyond reasonable doubt", the ideas of Interim Analyses and Stopping Rules are use.

Interim Analyses and Stopping Rules : The title should be, to some extent, self-explanatory. Interim Analyses are data analyses which are performed in advance of the scheduled end of a study. Pre-determined Stopping Rules are then applied to determine whether the trial should stop at that stage, due to the

emergence of overwhelming evidence in favour of a certain treatment, or whether it should continue further.

Clearly, if each interim analysis was performed at a significance level (S.L.) of 5%, the overall S.L. after several such analyses would be in excess of 5%. Armitage et al(1969) gave some resultant S.L.s for multiple testing at the 5% level. For example, even with only one interim analysis, the overall S.L. is increased to 8%, and with four interim analyses, this has risen to 14%.

Several different measures are possible to ensure that, overall, a 5% S.L. is preserved. For example, the S.L. for each analysis can be lowered by the same extent (see Pocock(1978)). This would mean that if, for example, a total of five analyses were performed, a result would only be considered to be significant if the observed significance was less than 0.016. Alternatively, it is possible to perform all interim analyses at a very stringent S.L., stopping the trial only for a very significant result, so that the final test could still be performed at just below the 5% level. (See Pocock(1983) for details).

Data Analysis : Not necessarily the most important job of the statistician, but often seen as such.

The analysis of clinical trials data is rarely straightforward, for a variety of reasons; in part due to the sheer volume of data generally collected, and also due to the wide ranges of possible approaches to handling such data. To highlight some of the problems, an example will be considered. Imagine that the aim is to "Compare the effects on blood pressure of two anti-hypertensive agents". Such a simple-sounding problem can disguise a fairly complicated task for the statistician :

- How will the "effects" be measured ? Will they be based on the raw blood pressure reduction over the study period, or on some dichotomisation of the results into blood pressure controlled / not controlled ?
- What if each patient visits the treatment centre for assessment at certain fixed periods during the study ? Should the blood pressures at these times also be incorporated into analyses using regression or repeated measures methodologies ?

- If intermediate results are to be incorporated, what if some of the planned visits are missed ?
- What should be done with the protocol violators ?
- Which prognostic variables should be incorporated into analysis of the responses of direct interest ?
- Is there any prior information available which could simplify analyses, or make them more efficient ?
- Should Normal or non-parametric techniques be used ?

The questions seem to be endless!

The multitude of problems behind such a simple question highlights the importance of having a clear set of objectives and a proposed method of analysis laid out in advance, in order to avoid accusations of "data-dredging" arising from having an ill-defined question of interest.

From the above section it is clear that there is a wide range of topics liable to be of interest to the practising medical statistician. To illustrate this, the following section aims to provide an overview of the current topics of interest, by reviewing some of the more recent issues of two leading medical statistics journals, Statistics in Medicine and Controlled Clinical Trials (years 1988 and 1989). Each article is classified according to its principal subject matter. It must be stated that the reviewing procedure used was purely subjective.

1.8 : Review of "Statistics in Medicine" and "Controlled Clinical Trials" 1988-1989

The results of reviewing "Statistics in Medicine" (SIM) and "Controlled Clinical Trials" (CCT) are shown in Table 1.1.

Since the style of these two journals is so different, with "Controlled Clinical Trials" inclined more towards the practical issues in performing clinical trials than "Statistics in Medicine", the reviews will be summarised separately.

Table 1.1 : Assessment of "Statistics in Medicine" and "Controlled Clinical Trials"

(Number of Papers)

RESEARCH TOPIC	SIM		CCT	
	1988	1989	1988	1989
STUDY DESIGN	4	3	2	3
SAMPLE SIZE / POWER CALCULATIONS	9	5	1	0
INTERIM ANALYSIS / STOPPING RULES	1	3	2	5
DATA ANALYSIS				
Repeated Testing / Multiple Comparisons	1	2	0	2
Meta-Analysis	1	1	0	3
Handling missing data	5	2	0	0
Ordinal Categorical data	2	4	0	0
Analysis of Covariance / Regression	2	6	0	1
Discrimination / Diagnosis	6	4	0	0
Identification of Clustering	3	1	0	0
Survival analysis	11	14	1	0
Case / Control Studies	2	5	0	0
Crossover Studies	1	2	0	1
Repeated Measures Studies	5	1	0	0
Time Series Analysis	4	5	0	0
Factorial Studies	0	2	0	1
General Statistical Modelling	7	3	0	0
EPIDEMIOLOGY				
Estimating trends	5	19	0	0
Assessment of risk factors	6	9	0	0
Estimation problems in epidemiology	2	4	0	0
Miscellaneous problems in epidemiology	4	5	0	0

(contd.)

Table 1.1 : contd.

MISCELLANEOUS				
Issues in Clinical Trials Methodology	4	7	14	10
Detailed Statistical Methodology	14	8	1	2
Medical Applications	6	6	0	0
CLINICAL TRIAL REVIEWS	0	0	1	7
TOTAL	105	121	22	35

1.8.1 : "Statistics in Medicine"

It can be seen that consistently, certain research topics received much attention compared to other important topics. The problems of survival analysis and the estimation of trends in epidemiology were devoted much attention, whilst, for example, the problems of stopping rules, the handling of missing data and the analysis of repeated measures/crossover studies received little emphasis, despite being of some considerable importance.

This leads to a concern that, perhaps, some of the areas requiring further investigation are being under-researched, while research could be being "duplicated" in the more popular areas.

It should be noted that this could conceivably be a slightly gloomy view of the situation : SIM periodically publishes special issues on topics of general interest. The January/February issue in 1988 looked at "Longitudinal Data Analysis" while the January 1989 issue looked at "Methods for Modelling the AIDS Epidemic", and so, probably, there will be disproportionately many papers on such topics as Estimating Epidemiological Trends. (In fact, $^{4}/_{5}$ of the papers in 1988 and $^{10}/_{19}$ of the papers in 1989 on that topic came from these special issues). Similarly, the March 1989 issue studied "Statistics in Surveillance", and so, again, it would be expected that there would be a high proportion of papers on epidemiological topics.

It could be reasoned, however, that these special issues would not have been produced if there had not been sufficient current interest in the topics covered.

1.8.2 : "Controlled Clinical Trials"

Certain features distinguish the style of CCT from that of SIM.

(i) CCT contains substantially fewer papers on detailed statistical methodology and analysis.

(ii) CCT tends to carry quite a number of papers purely reporting study findings.

(iii) CCT contains, mainly, a variety of general discussion papers on a wide range of topics relating to the planning and execution of clinical trials. These range from discussions on the role of the National Institutes of Health (N.I.H.) in the sponsorship of clinical trials

(1988 pp103-106) to the properties of randomisation (1988 No4 , Full Issue) to the problems of data collection (1989 pp282-289) and management (1989 pp386-406).

As with SIM, certain topics received a disproportionate amount of attention, these topics being interim analysis and stopping rules, while other, equally important issues such as meta-analysis and the handling of missing data received comparatively little attention.

1.9 : Conclusions for Chapter 1

The areas of research studied in this thesis were chosen to reflect a few of the "gaps" in the existing literature in clinical trials methodology.

This work began in 1985, but it is clear from the above review of the most recent two years of two leading medical statistics journals that these areas are still receiving little attention in the literature.

The first topic which will be studied is the problem of how best to handle incomplete data and/or dropouts in clinical trials, where inappropriate handling of such data can lead to substantial bias in the results obtained.

The second looks at a common problem in Phase II clinical trials : the analysis of data arising from treatment groups with an a priori ordering. Such data can be obtained if the responses of patients to differing doses of the same drug are being compared. Here, inappropriate analyses can lead to a substantial loss of statistical efficiency.

Both of these problems have been highlighted by the analysis of data from actual clinical trials, and indeed this research has been been sponsored by a pharmaceutical company (Janssen Pharmaceutica, Beerse, Belgium) which recognised the importance of developing appropriate statistical methodology to address these problems.

PART I : Inference in the Presence of Missing Data.

Chapter 2 : Background to the Problem

2.1 : Introduction and Literature Review

In the context of clinical trials, the problem of missing data is a very common one. Sometimes, even if a study protocol dictates that multiple measurements are made on each individual, not all of the possible measurements will be made. (This can happen either by design or by accident).

The situation where data are missing by design could arise when the time until the next visit of a patient to a blood pressure clinic depends on some critical threshold blood pressure, x, where if the blood pressure on a given visit exceeds x, the patient returns for assessment in two weeks, and if the blood pressure is less than x, the patient returns in four weeks.

The more common situation where observations are accidentally missing arises, for example, if a machine malfunction occurs (and no observation is taken), or if an individual forgets or is unable to return for assessment, or if unscheduled patient withdrawals are made due to, for example, side effects, etc.

The problem of missing data must, in fact, be as old as experimentation itself.

However, standard statistical procedures have, traditionally, been developed for the analysis of complete rectangular data sets (i.e the data may be laid out as a rectangle of values, the rows representing different cases and the columns representing different variables, with no "gaps" where observations should be).

Fortunately, over the years, several approaches to the treatment of missing data have been proposed, some of these on intuitive grounds, and others with more theoretical basis.

For any of the suggested approaches to give unbiased results, it is necessary to make certain assumptions about the mechanisms which could lead to data being missing. Several classifications of missing data mechanisms were defined by Rubin(1976) as explained in the following section.

2.2 : Missing Data Mechanisms

Let Θ be some parameter of the data of interest and let ϕ be the parameter of the process leading to the missing data. Then the missing data are defined to be Missing at Random (M.A.R.) if the conditional probability of the observed pattern of missing data given the missing data and the observed data is the same for all possible values of missing data (for each possible value of ϕ). This means that inferences about Θ can be based solely on the likelihood function for the observed data, without having to model the missing data mechanism. i.e "missingness" itself is uninformative. The observed data are defined to be Observed at Random (O.A.R.) if the conditional probability of the observed pattern of missing data given the missing data and the observed data is the same for all possible values of the observed data (for each possible value of ϕ and for each possible value the missing data). If the observed data are observed at random and the missing values are missing at random, then the missing data are said to be Missing Completely at Random. (See Rubin(1976) for details).

To illustrate these terms, consider a simple bivariate example.

Example : Let X and Y be two random variables, where X is completely observed and Y is incomplete, and let the data be represented as follows :

$$X_1\ X_2\ \ldots\ X_m\ X_{m+1}\ X_{m+2}\ \ldots\ X_n \quad \text{(i.e. X complete)}$$
$$Y_1\ Y_2\ \ldots\ Y_m \quad \text{(i.e. Y incomplete)}$$

Then $Y_{m+1}\ \ldots\ Y_n$ are missing at random if their probability of being missing depends on X but not on Y. The $Y_1\ \ldots\ Y_m$ are observed at random if their probability of being observed does not depend on X. Finally, $Y_{m+1}\ \ldots\ Y_n$ are missing completely at random if the probability of being missing is independent of X and Y.

Imagine, for example, the situation where any Y_j is only observed if the corresponding X_j takes a value lower than some predetermined threshold. Then, even although the Y_j are clearly missing in a systematic fashion, they are still, technically, Missing at Random, since their presence or absence does not depend on the value which they would have taken.

Methods described in the following sections will assume that,

at least, the data are missing at random, and often that they are also observed at random (i.e. missing completely at random). Although it would be theoretically possible to model the missing data mechanism (if it was known), this would seriously complicate any statistical analyses performed. In any case, generally no such information would be available. One common exception is where the data take the form of survival times. If a response is missing ("censored"), then that itself is informative - if a response is censored, then the implication is that the survival time exceeded the length of follow-up. In this situation of "informative missing data" it is certainly not reasonable to assume a "missing at random" scenario.

Since such censoring problems are well documented, and since standard analyses have long been available, no further discussion will be given.

The problems considered from now on will be those where no strong information is available about the mechanism leading to the missing data, or where it is known that the data are M.A.R.

The approaches to be considered fall into three principal categories :

(1) Relatively simple ad hoc approaches using available data (or a subset of the available data).

(2) Approaches where imputation is made for the missing values, and then complete-data techniques are employed.

(3) Approaches based on fitting certain models (usually Multivariate Normal) to the observed data and maximising the likelihood function obtained (valid when the missing data are missing at random).

In previous literature, the problems which have been addressed have, generally, been those of parameter estimation in isolated groups. In reality, however, it is more common to have two or more groups under study, where it is of interest to compare these groups in some sense, for example to test for equality of the group means while assuming a common covariance matrix, or to find interval estimates for differences between linear combinations of the group mean components, etc.

The following sections will, first, review and compare some of the possible approaches to parameter estimation in the single group situation, and then extend these approaches to cover the multiple group problem.

With the exception of those cases highlighted, the approaches which will be described are equally applicable in repeated measures applications (where each individual has repeated measurements made on the same variable through time) and in general multivariate applications.

2.3 : Approaches to Dealing with Missing Data

(1) Relatively Simple ad hoc Approaches

(a) Complete Cases (C.C.) : Analyses are performed using only those cases for which no data are missing. If the missing data are Missing Completely at Random, then this will be valid, leading to unbiased parameter estimation. However, there is the potential for the loss of important information by discarding the incomplete cases.

(b) All Available Data (A.A.D.) : Analyses are performed using all of the data on each variable, so that calculations for different variables will be based on observations from different numbers of individuals, making results, at best, difficult to interpret.

As long as the missing data are Missing Completely at Random, the estimates of means and variances will not be biased. However, problems arise when looking at combinations of two or more variables. To estimate covariances, clearly only cases with all of the variables of interest present may be used. Thus the estimates of variance of a variable and of its covariance with other variables may be based on different numbers of cases, leading to the possibility of estimates of correlation outside the range [-1, 1]. This problem can be solved by estimating the individual variances using only those cases with both variables of interest observed (so that the variance and covariance estimates are based on the same individuals). (See Matthai(1951) for details).

(2) Methods Based on Imputation

The idea of substituting in "suitable" values for missing variables is appealing in the sense that, after imputation has

been made, familiar complete-data methodology may be employed. The choice of method of imputation is largely subjective, with, as a result, a large number of options available. As in the previous sub-section, strong assumptions must be made about the missing data mechanisms in order that such analyses are valid.

(a) Imputing the Mean of The Observed Values For a Variable : For each variable under study, any missing values are replaced by the average of all available measurements on that variable. If a M.C.A.R. setup can be assumed, then clearly the sample means for the completed data matrix will give unbiased estimates for the true means (since the available case method did). However, such imputation will artificially reduce the estimated variances by a factor dependent on the number of cases where imputation was necessary.

(b) Last Value : For each variable under study, any missing values for a given case are replaced by the last observed value for that case. Clearly such an approach would only be sensible if the data represented repeated measurements on the same variable. For example, it would not make sense to impute an individual's height in place of their unknown weight.

Example : Measurements are intended to be made on the blood pressure of individuals at five defined time points. Imagine, instead, that the following data were observed (where the asterisks represent the missing values).

		Time Point				
		1	2	3	4	5
	1	170	165	122	*	*
	2	145	142	138	124	116
Case No.	3	129	118	*	115	115
	4	182	*	147	147	145
	5	131	135	*	*	*
	⋮					

Then if last value imputation was employed, the completed data matrix would take the form :

		Time Point				
		1	2	3	4	5
	1	170	165	122	**122**	**122**
	2	145	142	138	124	116
Case No.	3	129	118	**118**	115	115
	4	182	**182**	147	147	145
	5	131	135	**135**	**135**	**135**
	⋮					

(The values in bold type are those which would be imputed).

For this approach, even if the data were missing Completely at Random, the results obtained would be biased, unless the underlying means of the imputed variables were the same as those variables which were used to perform imputation - fairly unlikely in practice.

(c) Imputation using Regression Models : Represent the data as Y_{ik} where $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, p$,

with n = total number of cases

and p = number of variables which could be measured.

Then, if a Multivariate Normal model can be assumed for the data, much progress can be made.

Let $Y_i = (Y_{i1}, \ldots, Y_{ip})$ represent the data for case i.

Assume that $Y_i \sim N_p(\mu , \Sigma)$

Let μ_1 , Σ_1 represent the portions of μ and Σ , respectively, corresponding to the observed variables for case i, and let μ_2 , Σ_2 represent the portions corresponding to the missing values.

Let Σ_{12} represent the portion of Σ corresponding to covariances of combinations of one missing and one observed value for case i.

Then it would be intuitively appealling to replace the missing values for case i by their conditional expectations given the observed values for that case. If Y_1 are the observed values for case i and Y_2 are the missing variables, Y_2 would be imputed using regression on Y_1,

i.e Y_2 would be replaced by

$$\mathbb{E}(Y_2 \mid Y_1) = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (Y_1 - \mu_1)$$

where the components of μ and Σ would be estimated using

those cases with the relevant components present

Such imputation would be valid under M.C.A.R. as before, but would also be valid even if the missing data were only M.A.R. The means obtained under this method are unbiased for the true means, but the variances and covariances are, again, underestimated, this time due to failing to model any random variability about the regression model fitted.

Note : This method of imputation forms an important part of one of the likelihood - based methods to follow in the next section.

(3) Methods based on Modelling the Data

This third approach will be covered in detail in the following chapter. Briefly, the data are modelled under a M.A.R. assumption, using, most commonly, the Multivariate Normal distribution (although certain other models are possible, as will be seen later). Parameter estimates are then obtained using Maximum Likelihood techniques.

Chapter 3 : The Development of Likelihood-Based Approaches

3.1 : Introduction

This chapter first looks at various techniques for maximising the Multivariate Normal likelihood function for incomplete data from isolated groups. It then moves on to extend some of these techniques to cover the multiple group problem.

Also discussed are certain alternatives to the Multivariate Normal model, for which maximisation of the likelihood function can be achieved by only slight modifications of the techniques used for the Multivariate Normal model.

3.2 : The One-Sample Problem

3.2.1 : Introduction

Before proceeding to look at two general approaches for maximising Multivariate Normal likelihood functions in the presence of missing data, one special case will be considered

Throughout the following sections, it will be assumed that the missing data are Missing at Random (M.A.R.).

3.2.2 : Maximising the Likelihood Function When the Missing Data Form a Nested Pattern :

As long as they are missing at random, when the missing data form a nested pattern (nesting will be defined below), it can be shown that a useful factorisation of the Multivariate Normal likelihood function can be made, making parameter estimation straightforward.

Note : Different possible patterns of missing data can be simply described using a set of p binary variables, taking value 0 if the corresponding observation is absent and a value 1 if the corresponding observation is present.

For example, for the case p = 5,

[1 1 0 1 0] would represent the situation where the third and fifth measurements were missing, and the rest were observed.

Notation : Y_{ti} = Observed data for the i^{th} case with missing data pattern t ($i = 1, \ldots, n_t$; $t = 1, \ldots, T$)

For the complete cases, $Y_{ti} \sim N (\mu , \Sigma)$

For the incomplete cases, define a matrix, D_t, for each pattern, t, such that

$$Y_{ti} \sim N (\mu_t , \Sigma_t) = N (D_t \mu , D_t \Sigma D_t^T)$$

D_t is a matrix composed of rows each containing a single "1" with the remaining elements "0", such that each row picks out one observed component of the mean vector.

e.g. If $p = 5$, and the missing data pattern is [1 1 0 1 0], then the form of D_t would be

$$D_t = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The data Y_{ti} ($t = 1, \ldots, T$; $i = 1, \ldots, n_t$) are said to be nested if it is possible to label the groups so that the D_t ($t = 1, \ldots, T$) satisfy the condition

$$D_r = D_r D_k^T D_k \quad \text{for } r \succ k \qquad (3.1)$$

Putting this in more tangible terms, the types of data matrices satisfying (3.1) are, for example, those where the data can be arranged so that for a given missing data pattern, t+1, the observed variables are a subset of those measured in the previous pattern, t.

This would be the case if, for example, the data could be arranged into the patterns shown below :

	Pattern					
	1	1	1	1	1	Complete cases
	2	1	1	1	1	0 Cases with only last point missing
t	3	1	1	1	0	0 Cases with last 2 points missing
	4	1	1	0	0	0 Cases with last 3 points missing
	5	1	0	0	0	0 Cases with only first measurement

When the data form a nested structure, analyses under a Multivariate Normal model with M.A.R. assumed are relatively straightforward.

Return to the earlier notation for simplicity.

Let Y_{ij} = measurement on the j^{th} variable for case i.

$$(\ i = 1, \ldots, n \ ; \ \ j = 1, \ldots, p)$$

It can be seen that, for the example above, variable 1 is more observed than variable 2, which in turn is more observed than variable 3, etc. As a result, the corresponding likelihood can be factorised by dividing it into several components, each one being related to the conditional density of a variable, given all previous variables (see Anderson(1957)).

If we let $L(\ \Theta\ ;\ Y\ )$ represent the likelihood function,

$$L(\Theta;\ Y) = \prod_{i=1}^{n_1} p(Y_{i1}|\Theta_1) \prod_{i=1}^{n_2} p(Y_{i2}|Y_{i1},\Theta_2) \ldots \prod_{i=1}^{n_p} p(Y_{ip}|Y_{i1},..,Y_{i,p-1},\Theta_p)$$

where $\Theta_1,\ldots,\ \Theta_p$ are distinct, so that inferences about Θ_1 can be based on the n_1 components involving Θ_1 (i.e. those corresponding to cases with the first variable recorded), inferences about Θ_2 can be based on the n_2 components involving Θ_2 (i.e. those corresponding to cases with both the first and second variables recorded), etc.

Also, the joint density of $Y_1,\ \ldots,\ Y_p$ may be reconstructed using a similar expression, namely

$$p(Y_1,\ \ldots,\ Y_p|\ \Theta) = p(Y_1|\Theta_1)\ p(Y_2|Y_1,\Theta_2)\ \ldots\ p(Y_p|Y_1,\ \ldots Y_{p-1},\Theta_p)$$

Due to the factorisation of the likelihood function, maximum likelihood estimates for the parameters of the joint distribution can easily be found as functions of the p individual factors.

<u>Example</u> : $p = 2$; Bivariate Normal Model

$$Y_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} \sim N\left[\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right]$$

with the Y_{i1} complete and Y_{i2} incomplete.

$$P(Y_1,Y_2|\mu_1,\ \mu_2,\ \sigma_1,\ \sigma_2,\ \rho) = p(Y_2|Y_1,\ \alpha_{12},\ \beta_{12},\ \sigma_{12})\ p(Y_1|\mu_1,\ \sigma_1)$$

where α_{12} and β_{12} are the regression parameters for Y_{i2} on Y_{i1}, with

$$\begin{aligned} \alpha_{12} &= \mu_2 - \beta_{12}\ \mu_1 \\ \beta_{12} &= \rho\ \ \sigma_2 / \sigma_1 \\ \sigma_{12}^2 &= \text{Conditional variance of } Y_{i2} \text{ given } Y_{i1} \\ &= \sigma_2^2\ (1 - \rho^2) \end{aligned}$$

Let n_1 cases be complete and n_2 cases have only the second variable missing. The likelihood function is then

$$L(\theta\ ;\ Y) = \prod_{i=1}^{n_1} p(Y_{i1},Y_{i2}|\ \mu_1,\ \mu_2,\ \sigma_1^2,\ \sigma_2^2) \prod_{i=n_1+1}^{n_1+n_2} p(Y_{i1}|\ \mu_1,\ \sigma_1^2)$$

$$= \prod_{i=1}^{n_1} p(Y_{i2}|\ Y_{i1},\alpha_{12},\beta_{12},\sigma_{12}^2) \prod_{i=1}^{n_1+n_2} p(Y_{i1}|\ \mu_1,\ \sigma_1^2)$$

Maximum likelihood estimates for μ_1 and σ_1^2 are found by maximising the second factor. The results obtained are just the usual univariate Normal maximum likelihood estimates for mean and variance, namely

$$\hat{\mu}_1 = (n_1+ n_2)^{-1} \sum_{i=1}^{n_1+n_2} Y_{i1} = N^{-1} \sum_{i=1}^{N} Y_{i1} = \bar{Y}_1$$

where $N = n_1 + n_2$

$$\hat{\sigma}_1^2 = N^{-1} \sum_{i=1}^{N} (Y_{i1} - \bar{Y}_1)^2$$

To maximise the likelihood with respect to the conditional parameters, α_{12}, β_{12}, σ_{12}^2, only the first factor of the likelihood function need be considered. Here the usual regression estimates of the parameters are obtained, namely

$$\hat{\beta}_{12} = \frac{\sum_{i=1}^{n_1} (Y_{i1} - \bar{Y}_1')(Y_{i2} - \bar{Y}_2')}{\sum_{i=1}^{n_1} (Y_{i1} - \bar{Y}_1')^2}$$

$$\hat{\alpha}_{12} = \bar{Y}_2' - \hat{\beta}_{12}\bar{Y}_1'$$

$$\hat{\sigma}_{12}^2 = n_1^{-1} \left[\sum_{i=1}^{n_1} (Y_{i2}-\bar{Y}_2')^2 - \hat{\beta}_{12}^2 \sum_{i=1}^{n_1} (Y_{i1}-\bar{Y}_1')^2 \right]$$

where $\bar{Y}_1'$ and $\bar{Y}_2'$ are the averages of the Y_{i1} and the Y_{i2}, respectively, calculated using the n_1 complete cases. The outstanding parameters of the joint distribution are then found by substituting the maximum likelihood estimates for the conditional parameters into the formulae for μ_y, σ_y^2 and ρ as shown below.

$$\hat{\mu}_2 = \hat{\alpha}_{12} + \hat{\beta}_{12}\,\hat{\mu}_1$$

$$\hat{\sigma}_2^2 = \hat{\sigma}_{12}^2 + \hat{\beta}_{12}^2\,\hat{\sigma}_2^2$$

$$\hat{\rho} = \hat{\beta}_{12}\,\hat{\sigma}_1 / \hat{\sigma}_2$$

Clearly, as the number of variables (and the number of missing data patterns) increases, the calculations required to proceed from the parameters of the separate factors of the likelihood function to those for the joint distribution will become increasingly complex, but there would be no theoretical difficulties in so doing.

3.2.3 : Maximising the Likelihood Function when the Data are not Nested

More problematic is the situation where the data do not form a nested pattern (as defined in the previous section). In that case, no simple factorisation of the likelihood function is available to make calculation of the maximum likelihood estimates straightforward, and, in fact, solutions cannot be found analytically.

As a result, two main numerical techniques have been proposed in the past :

(i) The Method of Scoring

(ii) The EM Algorithm

(Note : Throughout the following sections a M.A.R. setup is assumed.)

These two numerical methods will be described in the context of a Multivariate Normal model for the data (although certain other distributions can be fitted easily using a minor modification of the latter technique).

(i) The Method of Scoring : This method (described by Hartley and Hocking(1971) for the situation where the data are Multivariate Normal and simplified by Hocking and Marx(1979))

involves solving the likelihood equations, i.e. solving the equations obtained when the first derivatives of the likelihood function with respect to μ and with respect to Σ are set to zero.

Consider the situation described earlier, where there are present T different missing data patterns. Then for each of these patterns, there will be defined a likelihood function - a Multivariate Normal likelihood based on the observed variables for cases with that particular pattern. The overall likelihood will then be given by the product of these T individual factors.

If we let L_t represent the likelihood function for pattern T and let L represent the total likelihood,

$$L = \prod_{t=1}^{T} L_t$$

with

$$\log(L_t) = l_t = (-n_t/2) \log |2\Pi\Sigma_t| - (1/2)\, tr(\Sigma_t^{-1} M_t)$$

where $$M_t = n_t (\hat{\Sigma}_t + (\hat{\mu}_t - \mu_t)(\hat{\mu}_t - \mu_t)^T)$$

$$\hat{\Sigma}_t = n_t^{-1} \sum_{i=1}^{n_t} (Y_{ti} - \hat{\mu}_t)(Y_{ti} - \hat{\mu}_t)^T$$

$$\hat{\mu}_t = n_t^{-1} \sum_{i=1}^{n_t} Y_{ti}$$

Thus, L can be seen to be a fairly awkward function to work with (and is unable to be factorised in the way shown in the previous section). It can be shown that the first derivative of the log-likelihood function with respect to μ is

$$\frac{\delta \log(L)}{\delta \mu} = - \sum_{t=1}^{T} n_t D_t^T \Sigma_t^{-1} D_t \mu - \sum_{t=1}^{T} n_t D_t^T \Sigma_t^{-1} \hat{\mu}_t$$

leading to likelihood equation for μ as

$$\sum_{t=1}^{T} n_t D_t^T \Sigma_t^{-1} D_t \mu = \sum_{t=1}^{T} n_t D_t^T \Sigma_t^{-1} \hat{\mu}_t$$

Similarly, the likelihood equations for Σ can be shown to be

$$\sum_{t=1}^{T} n_t D_t^T \Sigma_t^{-1} D_t = \sum_{t=1}^{T} D_t^T \Sigma_t^{-1} M_t \Sigma_t^{-1} D_t$$

Note here that the expression on the left is, in fact, the information matrix for μ (i.e. the second derivative of the log-likelihood function evaluated at the maximum likelihood estimates of the parameters) which gives a large-sample covariance matrix for the maximum likelihood estimate for μ.

Hocking and Marx(1979) made the estimation of μ and Σ relatively straightforward (compared to the original paper) by observing that the likelihood equations could be written as

$$\underline{\text{For } \mu} : \left(I - \sum_{t=2}^{T} (N_t/n_1) B_t D_t\right) \mu = \hat{\mu}_1 - \sum_{t=2}^{T} (N_t/n_1) B_t \hat{\mu}_t$$

$$\underline{\text{For } \Sigma} : \left(I - \sum_{t=2}^{T} (N_t/n_1) B_t D_t\right) \Sigma = n_1^{-1} M_1 + n_1^{-1} \sum_{t=2}^{T} (N_t/n_1)^2 B_t M_t B_t^T$$

where $N_t = \sum_{i=1}^{t} n_i$

n_1 = Number of cases in the complete block

and $B_t = (-n_t/N_t) \Sigma D_t^T \Sigma_t^{-1}$

This observation led to the suggestion of an efficient computational procedure for the maximum likelihood estimation of μ and Σ, as follows :

(1) Find initial estimates of the parameters μ and Σ (e.g. from the complete cases).
(2) Use the current estimate Σ to calculate B_t and hence μ.
(3) For the current estimate of μ, calculate M_t, and hence re-estimate Σ.
(4) Repeat steps (2) and (3) until successive estimates of the mean agree to within a specified tolerance.

For each iteration through (2) and (3), it can be seen that the p x p matrix in brackets on the left hand side of the likelihood equations need only be inverted once : the same expression is used in both steps (2) and (3). This is more efficient than in the original paper, where, for each

iteration, it was required to invert two such matrices - the information matrix for μ and that for Σ, where that for Σ was order $\frac{1}{2}p(p+1) \times \frac{1}{2}p(p+1)$.

As a direct consequence of the iterative procedure, easily obtained at the final step is the large - sample covariance matrix for the maximum likelihood estimate of μ, $\tilde{\mu}$ say. It can be shown that

$$\text{cov}(\tilde{\mu}) = (I - \sum_{t=2}^{T} (N_t/n_1)B_t D_t)^{-1} \Sigma/n_1 \qquad (3.2)$$

where n_1 is the number of cases in the complete block.

(ii) The EM Algorithm (see Dempster et al(1977)) : Let Y_{ij} represent the observation for case i on variable j ($i = 1, \ldots, n$; $j = 1, \ldots, p$).

As before, define partitions of the mean vector and covariance matrix according to the available data on a given case.

Let $Y_{1,i}$ represent the observed data in case i and let $Y_{2,i}$ represent the missing data in the same case.

The EM Algorithm involves the familiar elements of imputation for missing values, and the use of complete-data techniques. As with the Method of Scoring, initial parameter estimates are required. From that stage, the algorithm consists of two steps per iteration.

The Expectation Step (E - Step) : The missing values for a given case are replaced by their conditional expectations given the current parameter estimates and the observed data for that case.

The complete data sufficient statistics are then calculated using the completed data matrix at that iteration.

Let $Y_{ij}^{(t)}$ be the value of Y_{ij} at iteration t. Then

$$Y_{ij}^{(t)} = \begin{cases} Y_{ij} & \text{if } Y_{ij} \text{ observed} \\ \mathbb{E}(Y_{ij} | Y_{1,i}, \theta^{(t)}) = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}(Y_{1,i} - \mu_1) & \text{if } Y_{ij} \text{ missing} \end{cases}$$

where $\theta^{(t)}$ represents the parameter estimates at iteration t.

The sufficient statistics then calculated (in order to be able to estimate μ and Σ at that iteration) are as shown below

(a) S_y , a p-dimensional vector with j^{th} component

$$S_y(j) = \mathbb{E}\left(\sum_{i=1}^{n} Y_{ij} \mid Y_{1,i}, \theta^{(t)} \right) = \sum_{i=1}^{n} Y_{ij}^{(t)}$$

(b) S_{yy} , a p x p matrix with $(j,k)^{th}$ element

$$S_{yy}(j,k) = \mathbb{E}\left(\sum_{i=1}^{n} Y_{ij}Y_{ik} \mid Y_{1,i}, \theta^{(t)} \right)$$

$$= \sum_{i=1}^{n} Y_{ij}^{(t)}Y_{ik}^{(t)} + c_{jki}^{(t)}$$

where $c_{jki}^{(t)}$ is an estimate at iteration t of $\mathrm{cov}(Y_{ij},Y_{ik} \mid Y_{1,i}, \theta^{(t)})$, given by

$$c_{jki}^{(t)} = \begin{cases} 0 & \text{if either } Y_{ij} \text{ or } Y_{ik} \text{ observed} \\ (j,k)^{th} \text{ element of } \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} & \\ & \text{otherwise} \end{cases}$$

The Maximisation Step (M - Step) : This step is straightforward. The estimates of μ and Σ are updated using the various summary measures of the completed data matrix shown above :

j^{th} element of μ :

$$\mu_j^{(t+1)} = n^{-1} \sum_{i=1}^{n} Y_{ij}^{(t)} \qquad j = 1, \ldots, p$$

$$= n^{-1} S_y(j)$$

$(j,k)^{th}$ element of Σ :

$$\Sigma_{jk}^{(t+1)} = n^{-1} \mathbb{E}\left(\sum_{i=1}^{n} Y_{ij}Y_{ik} \mid Y_{1,i}, \theta^{(t)} \right) - \mu_j^{(t+1)}\mu_k^{(t+1)}$$

$$= n^{-1} \left(S_{yy}(j,k) - n^{-1} S_y(j) S_y(k) \right)$$

This algorithm proceeds by alternation between E - Step and the M - Step until successive parameter estimates agree to within some specified tolerance.

Note : This algorithm is not equivalent to filling in conditional expectations for the missing data and then using complete data techniques. The calculation of the sufficient

statistics at each stage explicitly takes into account the fact that certain of the data were originally missing.

3.2.4 : Discussion

Since both the EM Algorithm and the Method of Scoring maximise the likelihood function, the parameter estimates obtained under the two methods will be identical.

In the discussion section of Dempster, Laird and Rubin(1977) Murray pointed out that there can sometimes be problems of multiple maxima in the likelihood function, and, dependent on the starting values for the parameters, the EM Algorithm will sometimes converge to a local, rather than the global, maximum. This problem can be solved by running the algorithm for several different starting values and comparing the results.

Three possible options for the starting values are :

(i) μ set to zero and Σ set to be an identity matrix

(ii) μ and Σ set to be their complete-case estimates

(iii) μ and Σ estimated using complete-data techniques after some form of simple imputation has been made for the missing values.

So, which of the two numerical methods, described above, should be used ?

Two arguments often put forward to recommend the Method of Scoring as opposed to the EM Algorithm are :

(i) The large-sample covariance matrix for the maximum likelihood estimate of μ is given out automatically at the end of the Method of Scoring.

(ii) The convergence rate for the Method of Scoring is greater than for the EM Algorithm (the rate being quadratic rather than linear as in the EM approach).

These features of the Method of Scoring do not give cause to "write off" the EM Algorithm in any sense. Feature (i) above is not such an immense advantage as is sometimes conveyed in the literature. The large-sample covariance matrix for the maximum likelihood estimate of μ can easily be found at the end of the EM Algorithm by substituting the estimates obtained at the EM Algorithm's final step into equation (3.2) in the Method of

Scoring section, i.e. all that is required is one extra matrix inversion at the end of running the EM Algorithm - hardly a severe problem.

Feature (ii) above cannot be dismissed so lightly. The convergence rate for the Method of Scoring is certainly superior to that of the EM Approach. However, with powerful computers now readily accessible, convergence rate is not such a serious issue as it was in the past. Certainly, in the analysis of a single data set, there would be little difference in the times required for running the two algorithms. If, however, a large-scale simulation study was to be performed, the comparative convergence rates would be a more important issue.

Some features to recommend the EM Algorithm as opposed to the Method of Scoring are :

(i) The EM Algorithm is very simple to implement and is appealing on intuitive grounds - it calls upon "natural" techniques which have long been used, and yet has a firm theoretical basis.

(ii) Because of its simplicity, the algorithm is not difficult to amend to perform other, more specialised, tasks. e.g. fitting models for distributions other than Multivariate Normal, fitting multiple groups with a common covariance matrix, etc.

So far, several different approaches for dealing with missing data have been described, each approach reliant on certain assumptions about the missing data mechanism. The approaches have ranged from several ad hoc and intuitively appealing approaches (sometimes, but not always, leading to unbiased results in the presence of a M.C.A.R. mechanism), to approaches with more theoretical grounding, based on maximising the likelihood function, after making distributional assumptions about the data themselves. The latter were only valid if the data were Missing at Random. Three alternative numerical methods were described for maximising the likelihood function.

3.3 : The K-Sample Problem

3.3.1 : Introduction

As mentioned earlier, although it is useful to have at one's disposal valid approaches for dealing with single groups in isolation, it is much more common to have several groups under study, where it is of interest to compare these groups in some sense. It is in such areas that the remaining work will be concentrated.

Note : Since there exist standard approaches for dealing with complete Multivariate Normal data, there will be no computational difficulties in performing the ad hoc analyses or analyses using imputed data as before.

3.3.2 : Fitting the Model $\mu_1, \mu_2, \ldots, \mu_K, \Sigma$ to Incomplete K - Sample Multivariate Normal Data

The aim is to fit a "common covariance matrix, separate means" MVN model to data consisting of multivariate measurements on N individuals divided into K separate groups. The algorithm proposed below for this situation shows more than a passing similarity to the EM Algorithm described earlier for the one - sample problem. As a result, the previous algorithm will now be referred to as EM(a), while the following algorithm will be referred to as EM(b).

Notation : Let Y_{gij} represent the observation for case i in the g^{th} group on variable j ($i = 1, \ldots, n_g$; $g = 1, \ldots, K$; $j = 1, \ldots, p$. For each case, let $Y_{gi} = (Y_{gi1}, \ldots, Y_{gip})$, and define a partition of their own group's mean vector μ_g and the covariance matrix, Σ, as follows :

$$\mu_g = \begin{bmatrix} \mu_{1,i}(g) \\ \mu_{2,i}(g) \end{bmatrix} \qquad \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where $\mu_{1,i}(g)$ contains the elements of μ_g corresponding to the observed variables in case i, and $\mu_{2,i}(g)$ contains those corresponding to the missing variables in case i. Similarly, Σ_{11} contains the elements of Σ corresponding to the observed values,

Σ_{22} contains those for the missing values and $\Sigma_{21} = \Sigma_{12}^T$ contains those for combinations of observed and missing variables.

The Method EM(b)

(1) Start with initial estimates of the K mean vectors, and the common covariance matrix (e.g. from the complete cases).

(2) Complete the data matrix using

$$Y_{mij}(t) = \begin{cases} Y_{mij}^{(t)} & \text{if } Y_{mij} \text{ observed} \\ \mathbb{E}(Y_{mij} \mid Y_{1,i}, \theta^{(t)}) = \mu_{2,i}(m) - \Sigma_{21}\Sigma_{11}^{-1}(Y_{1,i} - \mu_{1,i}(m)) & \text{if } Y_{mij} \text{ missing} \end{cases}$$

where $Y_{1,i}$ and $Y_{2,i}$ represent, respectively, the observed values and the missing values for case i.

Calculate sufficient statistics for the completed data matrix, these statistics being :

$$\mathbb{E}\left(\sum_{i=1}^{n_m} Y_{mij} \mid Y_{1,i}, \theta^{(t)}\right) = \sum_{i=1}^{n_m} Y_{mij}^{(\epsilon)}$$

for $j = 1, \ldots, p$; $m = 1, \ldots, K$

$$\mathbb{E}\left(\sum_{i=1}^{n_m} Y_{mij}Y_{mik} \mid Y_{1,i}, \theta^{(t)}\right) = \sum_{i=1}^{n_m} Y_{mij}(t)Y_{mik}(t) + c_{jki}$$

where $c_{jki} = \text{cov}(Y_{mij}, Y_{mik} \mid Y_{1,i}, \theta^{(t)})$

$$= \begin{cases} 0 & \text{if either variable observed} \\ (j,k)^{\text{th}} \text{ element of } \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} & \text{if both variables missing} \end{cases}$$

(3) Calculate updated estimates of the K mean vectors and the common covariance matrix using the summary statistics from (2) above as follows :

The j^{th} Component of μ_m :

$$\mu_j(m)^{(t+1)} = n_m^{-1} \sum_{i=1}^{n_m} Y_{mij}(t)$$

where n_m = Number of cases in group m.

The $(j,k)^{th}$ element of Σ :

$$\Sigma_{jk}^{(t+1)} = N^{-1} \sum_{m=1}^{K} \{ \mathbb{E}(\sum_{i=1}^{n_m} Y_{mij} Y_{mik} | Y_{1,i}, \Theta^{(t)}) - n_m \mu_j(m)^{(t+1)} \mu_k(m)^{(t+1)} \}$$

(4) Repeat steps (2) and (3) until successive estimates of the mean vectors agree to within some specified tolerance.

The use of algorithm EM(b) easily leads to calculation of the maximised likelihood for the problem under consideration.

3.3.3 : A Likelihood Ratio Testing Procedure for K - Sample Multivariate Normal Data

In the situation where we have complete K-group multivariate data, a common testing procedure carried out is as follows (where each test performed is a likelihood ratio test).

Stage 1 : Test for equality of the K covariance matrices,

i.e. test

$$H_0 : \Sigma_1 = \ldots = \Sigma_K = \Sigma \; ; \mu_i \quad \text{vs} \quad H_1 : \mu_i \; ; \Sigma_i$$

i.e.

$$H_0 : Y_{mij} \sim N_p(\mu_m, \Sigma) \quad \text{vs} \quad H_1 : Y_{mij} \sim N_p(\mu_m , \Sigma_m)$$

If H_0 may not be rejected, go on to stage 2.

Stage 2 : Test for equality of the K group means under an assumption of equality of the covariance matrices,

i.e. test

$$H_0 : \mu_1 = \ldots = \mu_K = \mu \; ; \Sigma \quad \text{vs} \quad H_1 : \mu_i \; ; \Sigma$$

i.e. test

$$H_0 : Y_{mij} \sim N_p(\mu , \Sigma) \quad \text{vs} \quad H_1 : Y_{mij} \sim N_p(\mu_m , \Sigma)$$

In the tests described at Stage 1 and Stage 2 above, the test statistic used is the likelihood ratio test statistic,

$$2 \log(\lambda) = 2(l_1 - l_0) \sim \chi^2(\dim H_1 - \dim H_0) \text{ under } H_0$$

Here,

l_0 = maximised log-likelihood under the null hypothesis being tested ,

l_1 = maximised log-likelihood under the alternative

hypothesis of interest ,

dim H_0 = Number of fitted parameters under H_0 ,

dim H_1 = Number of fitted parameters under H_1 .

The algorithm EM(b) leads to similar testing procedures being possible in the incomplete data situation. The required maximised log-likelihoods are calculated as follows :

(a) For model (μ , Σ), all of the available data are pooled together, and EM(a) is run, treating the data as though from a single group.

(b) For model (μ_1, ..., μ_K, Σ), algorithm EM(b) is run.

(c) For model (μ_1, ..., μ_K; Σ_1, ..., Σ_K), algorithm EM(a) is run on the K groups separately, giving K groupwise maximised log-likelihoods. The overall maximised log-likelihood is obtained by summing up these K individual quantities.

The likelihood-ratio testing procedure described at Stages 1 and 2 above can then be performed using these maximised log-likelihoods from the incomplete data. Thus, as desired, the complete data procedures can be emulated in the incomplete data situation.

NOTE : It is well documented that covariance tests, for example Bartlett's Test and its Multivariate analogue, tend to be sensitive to non-Normality of the underlying data. Because of this, any significant results from Stage 1 should be tempered with the appropriateness, or otherwise, of the Normal model for the data. In certain cases, a significant result could be more indicative of an inappropriate Normal model than of a real difference between the groups.

3.3.4 : Follow-Up Procedures

Having rejected the null hypothesis at Stage 2, it would be desirable to perform the same sorts of follow-up procedures that would be carried out if the data were complete, for example

(i) Find out which of the K means are significantly different.

(ii) Find interval estimates for differences between linear combinations of the mean vector components for the

different groups

e.g. In a hypertension trial measuring blood pressure over a number of hospital visits while concurrently administering one of a possible set of anti-hypertensive agents, of interest might be to compare the changes in blood pressure produced during the study by the different agents.

Large Sample Covariance Matrices for the Maximum Likelihood Estimates of the K Means

To perform either of these tasks, it is necessary to produce some kind of estimated covariance matrices for the maximum likelihood estimates of the K mean vectors under the model $\mu_1, \ldots, \mu_K, \Sigma$.

Utilising the methods of Hocking and Marx(1979) in the one-sample problem, it can be shown that the large-sample covariance matrix for the maximum likelihood estimate of the k^{th} group mean is given by

$$\operatorname{cov}(\hat{\mu}_k) = [\sum_{t=1}^{T_k} n_{tk} D_t^T \Sigma_t^{-1} D_t]^{-1}$$

where D_t , Σ_t are as defined in the one-sample problem,

n_{tk} = number of cases in group k with missing data pattern t ,

and T_k = number of different "patterns" observed in group k

while $\operatorname{cov}(\hat{\mu}_k, \hat{\mu}_s) = 0$ for $k \neq s$.

Approximate 95% Confidence Intervals

Returning to the questions of interest

Problem (i) : Of interest in comparing the K groups would be to find intervals for expressions of form $\sum_{i=1}^{K} c_i \mu_i$. Clearly this expression will have a maximum likelihood estimate of $\sum_{i=1}^{K} c_i \hat{\mu}_i$

and an estimated covariance matrix of $\sum_{i=1}^{K} c_i^2 \hat{cov}(\hat{\mu}_i)$.

This leads to approximate 95% confidence intervals for the expressions of interest of form

$$\left\{\mu_i : \left(\sum_{i=1}^{K} c_i(\hat{\mu}_i-\mu_i)\right)^T \left(\sum_{i=1}^{K} c_i^2\, \hat{cov}(\hat{\mu}_i)\right)^{-1} \left(\sum_{i=1}^{K} c_i(\hat{\mu}_i-\mu_i)\right) < \chi^2(p;\ 0.95)\right\}$$

where p = number of repeated measures.

Clearly, if several such intervals were required (e.g. if all of the pairwise group comparisons were to be made) then some allowance for the multiple comparisons would have to be made in the χ^2 deviate chosen.

Problem (ii) : Of interest was to look at differences between linear combinations of the mean vector components in the different groups, that is, look at expressions of the form

$$c^T\mu_i - c^T\mu_j .$$

The maximum likelihood estimate of this expression would simply be

$$c^T\hat{\mu}_i - c^T\hat{\mu}_j = c^T(\hat{\mu}_i - \hat{\mu}_j)$$

with corresponding large-sample covariance matrix given by :

$$c^T(\hat{cov}(\hat{\mu}_i) + \hat{cov}(\hat{\mu}_j))\, c ,$$

leading to approximate 95% confidence intervals of form :

$$c^T(\hat{\mu}_i - \hat{\mu}_j) \pm N(0, 1; 0.975)\ \sqrt{\{c^T[\hat{cov}(\hat{\mu}_i) + \hat{cov}(\hat{\mu}_j)]\, c\}}$$

Again, clearly, if several such intervals were to be calculated, some allowance would be required for the multiple-comparisons aspect of the problem.

Alluded to earlier was that the EM Algorithm may be modified to fit models other than Multivariate Normal. These other models tend to be aiming, in a sense, to downweight the more "extreme" data values, and so could be thought of as robust methods.

3.4 : Robust Approaches

3.4.1 : Introduction

The use of robust methods was clearly described in Little(1988), where the models considered were the Multivariate t distribution (which could be thought of as similar to the Multivariate Normal distribution, but with more weight concentrated in the "tails"), and the Contaminated Multivariate Normal distribution (where one Multivariate Normal distribution is contaminated by forming a mixture distribution, with a small weight being associated with a much more diffuse Multivariate Normal distribution). In each of these models, weights are assigned to cases, the more extreme points being assigned lower weights. What would be desired would be that :

(i) For the Multivariate t (MVt) model, a steadily decreasing set of weights would be assigned as the data became more extreme*.

(ii) For the Contaminated Multivariate Normal (CN) model, similar weights would be assigned to the majority of cases with very low weights being given only to those cases deemed to be "contaminating".

* The classification of cases into levels of "extremity" is performed using the squared Mahalanobis distances for the cases, calculated using their observed data, i.e.

$$d_i^2 = (Y_{1,i} - \mu_1)^T \Sigma_{11}^{-1} (Y_{1,i} - \mu_1)$$

where μ_1 and Σ_{11} are calculated using the EM Algorithm.

To carry out the robust estimation procedures, the general form of model is as given below (which covers the MVN model as a special case).

Model : Conditional on unknown scalars $q_1, \ldots, q_n$, $Y_1, \ldots, Y_n$ form a random sample from a Multivariate Normal Distribution , $N (\mu , (1/q_i)\Sigma)$, where the q_i are independent and identically distributed, from distribution $h(q_i)$, so that the form of $h(q_i)$

will determine the distribution of Y.

Then it is possible to find maximum likelihood estimates of μ and Σ by treating the $Y_{2,i}$ and the q_i as missing data.

In this situation, if there were no missing data, then the data would be from a regular exponential family, with sufficient statistics

$$S_0 = \sum_{i=1}^{n} q_i \qquad S_y = \sum_{i=1}^{n} q_i Y_i \qquad S_{yy} = \sum_{i=1}^{n} q_i Y_i Y_i^T$$

leading to maximum likelihood estimates of μ and Σ as

$$\hat{\mu} = \sum_{i=1}^{n} (q_i Y_i) \Big/ \sum_{i=1}^{n} q_i = S_y / S_0$$

$$\hat{\Sigma} = n^{-1} (S_{yy} - S_y S_y^T / S_0)$$

Thus if there are missing data, the extension of the standard EM Algorithm is obvious :

(1) Estimate S_0, S_y and S_{yy} by their conditional expectations given the observed data and the current parameter estimates.

(2) Re-estimate μ and Σ using S_0, S_y and S_{yy}.

(3) Repeat steps (1) and (2) until successive parameter estimates agree to within a given tolerance.

The Conditional Expectations of S_0, S_y and S_{yy}

S_0 : $$\mathbb{E}(S_0 | Y_{1,i}) = \mathbb{E}\left(\sum_{i=1}^{n} q_i | Y_{1,i}\right) = \sum_{i=1}^{n} w_i^{(t)}$$

where $w_i^{(t)}$ is the estimate for q_i at iteration t, the form of $w_i^{(t)}$ dependent on the model being fitted.

The j^{th} Component of S_y :

$$\mathbb{E}(S_y(j) | Y_{1,i}) = \sum_{i=1}^{n} \mathbb{E}[q_i \, \mathbb{E}(Y_{ij} | Y_{1,i}, q_i) \mid Y_{1,i})$$

$$= \sum_{i=1}^{n} w_i^{(t)} \, Y_{ij}^{(t)}$$

where $Y_{ij}^{(t)} = \mathbb{E}(Y_{ij} | Y_{1,i})$ as in the standard EM Algorithm.

The $(j,k)^{th}$ Component of S_{yy} :

$$\begin{aligned}
&\mathbb{E}(S_{yy}(j,k) \mid Y_{1,i}) \\
&= \mathbb{E}\left(\sum_{i=1}^{n} q_i Y_i Y_i^T \mid Y_{1,i}\right) \\
&= \sum_{i=1}^{n} \mathbb{E}\left(q_i \, \mathbb{E}(Y_i Y_i^T \mid Y_{1,i}, q_i) \mid Y_{1,i}\right) \\
&= \sum_{i=1}^{n} \mathbb{E}\left(q_i (Y_{ij}(t) Y_{ik}(t) + \mathrm{cov}(Y_{ij}(t), Y_{ik}(t) \mid Y_{1,i}, q_i) \mid Y_{1,i}\right) \\
&= \sum_{i=1}^{n} w_i(t)\, Y_{ij}(t)\, Y_{ik}(t) + c_{jki}
\end{aligned}$$

where

$$c_{jki} = \begin{cases} 0 & \text{if } Y_{ij} \text{ or } Y_{ik} \text{ observed} \\ (j,k)^{th} \text{ element of } \Sigma_{22}(t) - \Sigma_{21}(t)\Sigma_{11}^{-1}(t)\Sigma_{12}(t) & \text{otherwise} \end{cases}$$

All that remains is to specify the form of $h(q_i)$, and hence the weights w_i. As mentioned earlier, the form of $h(q_i)$ and hence w_i depends on the model being fitted.

3.4.2 : The Models

The Multivariate t Model (MVt) : It can be shown that to achieve a distribution for Y_i as $t_p(\mu, \Sigma, \nu)$, the suitable choice of model for q_i is that $q_i\nu \sim \chi^2(\nu)$. If that is the case, the required case weights are

$$\begin{aligned}
w_i(t) &= \mathbb{E}(q_i \mid Y_{1,i}, \mu(t), \Sigma(t)) \\
&= (\nu + p_i) / (\nu + d_i(t)^2)
\end{aligned}$$

where $d_i(t)^2$ is the estimate of d_i^2 at iteration t and p_i is the number of variables observed in case i (since the distribution of $(\nu + d_i^2)q_i$ conditioned on $Y_{1,i}$ and the current parameter estimates is $\chi^2(\nu + p_i)$).

The Contaminated Multivariate Normal Model (CN) : If it is desired that Y_i be Contaminated Multivariate Normal (i.e a

mixture of two Multivariate Normal distributions), $N(\mu, \Sigma)$ in proportion $(1-\delta)$ and $N(\mu, \Sigma/\lambda)$ in proportion δ, where δ (the "contaminating fraction") is very small and $0 < \lambda < 1$ (δ and λ both known), then the appropriate form of $h(q_i)$ is

$$h(q_i) = \begin{cases} 1 - \delta & \text{if } q_i = 1 \\ \delta & \text{if } q_i = \lambda \\ 0 & \text{otherwise} \end{cases}$$

For this model, the correct form for the case weights is

$$w_i(t) = \frac{1 - \delta + \delta\lambda^{1+0.5pi} \exp\{0.5(1-\lambda)d_i(t)2\}}{1 - \delta + \delta\lambda^{0.5pi} \exp\{0.5(1-\lambda)d_i(t)2\}}$$

where $d_i(t)2$ the estimate of d_i^2 at iteration t.

Incidentally, to fit a Multivariate Normal (MVN) model, unit weights are assigned to all cases.

For each of the models described (MVN, MVt, CN), clearly it would not be difficult to calculate the likelihood function corresponding to the data at each iteration.

For a MVN model, the likelihood function is as given earlier.

For a MVt model, the log-likelihood function is defined by :

$$l = \sum_{i=1}^{n} [\, -0.5 \log|2\Pi\Sigma_{11}| - 0.5(\nu + p_i) \log(1+d_i^2/\nu)$$
$$-0.5p_i\log(0.5\nu) + \log\{\, \Gamma(0.5(\nu+p_i) / \Gamma(0.5\nu) \,\} \,]$$

while for a CN model, the log-likelihood function is given by

$$l = \sum_{i=1}^{n} [\, -0.5 \log|2\Pi\Sigma_{11}| - 0.5\, d_i^2$$
$$+ \log(1 - \delta + \delta\lambda^{0.5pi} \exp(0.5\,(1-\lambda)\, d_i^2) \,]$$

where p_i = number of variables observed for case i.

It would then seem natural to assess the relative merits of the three models by comparing their maximised log-likelihoods.

3.4.3 : The Choice of Model

Superficially, it would be desired that in many circumstances,

the Multivariate Normal model would be sufficiently flexible to provide reasonable fit to the data under study (after transformation of the data if necessary), with more complicated models only used as a "last resort".

So, a useful initial step would be to assess, in some way, the validity of the MVN model. One method described by Gnanadesikan (1977), Little and Smith(1987) and Little(1988) involved the d_i^2 quantities defined earlier, and is known as the "Transformed Distance Plot".

The Transformed Distance Plot

If a Multivariate Normal model was correct for a given data set, then the squared Mahalanobis distances ($\hat{d}_i^2$) obtained at the final iteration of the MVN EM Algorithm would have $\chi^2(p_i)$ distribution (where p_i is the number of variables observed for case i). These $\hat{d}_i^2$ could then be transformed to approximately Standard Normal deviates by using the so-called "Wilson Hilferty Transformation" as outlined in Johnson and Kotz(1970), defined as

$$P(\chi^2(\nu) < x) \simeq \Phi \left[(x/\nu)^{1/3} - 1 + 2/9\,\nu^{-1} \right] \sqrt{(9\nu/2)}$$

where Φ represents the Standard Normal cumulative distribution function.

Using this result, it can be seen that if the data were indeed Multivarite Normal,

$z_i = \{ (d_i^2/p_i)^{1/3} - 1 + 2/(9p_i) \} \sqrt{(9p_i/2)}$ would have an approximate N(0,1) distribution, while unusually large values of z_i would suggest that the data were more dispersed than would be expected for a Multivariate Normal distribution. In the examples to follow later, a very simple technique was used to assess objectively whether observed values of z_i were "too large" or were what could be reasonably expected within a MVN framework.

The Plots.

Let n be the sample size. Nineteen N(0,1) order statistics of size n were generated using NAG program G05DDF. Then for each of the ordered z_i from the data, the range of the corresponding twenty simulated values was calculated. This gave an "envelope" within which the observed z_i could be expected to lie if a MVN

model was reasonable, the "envelope" composed of the n pointwise intervals. It could be expected that if the data were MVN, that the z_i would fall ouside the "envelope" approximately $n/20$ times. (See Barnard(1963), Marriott(1979) for details).

In any examples given later, the form of the transformed distance plots will be with :

(i) z_i plotted against $\mathbb{E}(z_i)$ (the expected value of the corresponding point in a N(0,1) order statistic. Letting $z_{(i)}$ denote the i^{th} smallest z-value,

$$\mathbb{E}(z_{(i)}) = \Phi^{-1}[(i- 3/8)/ (n+1/4)] \quad .$$

(ii) A dotted line drawn in at $z_i = \mathbb{E}(z_i)$.

(iii) The pointwise intervals joined to give a continuous "envelope".

If such a plot contained many more points outside the intervals than could be expected by chance, then it would be necessary to look for some alternative way to model the data under study, for example using the robust methods already described. Even specifying a particular form of robust model still leaves some important choices to be made :

If a MVt model is chosen : How many degrees of freedom would be appropriate to give good fit to the data (i.e. the choice of parameter ν) ?

If a CN model is chosen : What contamination fraction and variance inflation should be used (i.e. the choice of δ and λ) ?

A natural approach would be to use the parameter values which maximised the likelihood function. This is not difficult for the MVt model, where the maximised likelihoods for various values of ν could be plotted , and the overall maximising value of ν found by inspection. The choice of the two parameters, δ and λ, for the CN model is more awkward. In theory it would be possible to calculate the maximised likelihoods over a grid of values for δ and λ and pick out the maximising combination. In practice, this would involve much effort. More simple would be if either of these parameters was known in advance, so that the maximisation was only necessary over one parameter rather than two.

To decide objectively whether a given robust model fits significantly better than the MVN model, their maximised log-likelihoods can be compared. However, only for the MVN vs MVt comparison are standard likelihood ratio tests valid. In the case

where the MVN and CN models are to be compared there is failure of the regularity conditions on which the asymptotic theory of the likelihood ratio test is based.

The one- and K- Sample procedures covered so-far will now be illustrated by way of several examples. Throughout, the notation for representing missing data patterns given in the "nested data" section will be used.

Chapter 4 : A Comparative Study Based on Three Clinical Data Sets

4.1 : Ketanserin vs Metoprolol in the Treatment of Hypertension.

(see Rosendorff & Murray(1986) ; Murray & Findlay(1988); Lewis(1989))

Background : A multicentre randomised clinical trial was performed to assess and compare the hypotensive (blood pressure lowering) effects of two drugs, Metoprolol (a β-blocker) and Ketanserin (an S_2 receptor blocking agent) during a twelve-week study period.

Assessment of individuals was carried out at five stages : Week 0 (Randomisation) , plus Weeks 2, 4, 8 and 12 weeks after commencement of active treatment. In addition, there followed an "open-phase" where patients with inadequate blood pressure reduction in the preceding period could be administered additional compounds in order to reduce the blood pressure to a more satisfactory level (The open phase will not be analysed here).

In the study protocol, it was stated that patients with diastolic blood pressure exceeding 110 mmHg at either the four-week or the eight-week assessment could "jump" straight to the open phase of the study, due to ethical problems associated with maintaining patients on a treatment which was, for them, ineffective.

Thus, clearly, it will be expected that there will be several patients with missing data at the end of the study, and so provision for such cases must be made in analyses performed.

The Data : The data consisted of 429 cases, split into 211 Ketanserin cases and 218 Metoprolol cases, with missing data patterns as shown in Table 4.1. For both drugs, the majority of cases with missing data were those who had "jumped" to the open phase. Since, in these cases, the reason for the data being missing was related only to the earlier blood pressure measurements, the missing data could be considered to be Missing at Random in the sense of Rubin(1976), although they could

Table 4.1 : Missing Data Patterns

Ketanserin Group

Weeks after Randomisation					Number of Cases
0	2	4	8	12	
1	1	1	1	1	136
1	1	1	1	0	19
1	1	1	0	0	41
1	1	0	0	0	4
1	0	0	0	0	1
1	1	1	0	1	1
1	1	0	1	1	2
1	0	1	1	1	2
1	0	1	1	0	1
1	0	1	0	1	1
1	0	0	1	1	2
1	0	0	1	0	1
					211

Metoprolol Group

Weeks after Randomisation					Number of Cases
0	2	4	8	12	
1	1	1	1	1	162
1	1	1	1	0	14
1	1	1	0	0	30
1	1	0	0	0	2
1	0	0	0	0	1
1	1	1	0	1	2
1	1	0	1	1	1
1	0	1	1	1	2
1	0	1	0	1	1
1	0	1	0	0	1
1	0	0	1	1	1
1	0	0	0	1	1
					218

certainly not be considered to be Missing Completely at Random - In no way could the incomplete cases be described as a random sample of all the cases under study.

In addition to the planned withdrawals during active treatment, there were also several other "unplanned" cases of missing data. For simplicity in analyses, due to the small proportion of unplanned missing data, these cases will also be assumed to be missing at random, although this will not necessarily be the case.

Noticeable from the observed missing data patterns is that there are many more planned withdrawals in the Ketanserin group than in the Metoprolol group, suggesting that the Ketanserin was not as successful in controlling blood pressure quickly as the Metoprolol.

The Analyses

Comparison of the EM Algorithm Results to those from Three Naive Approaches :

For each group separately, the five-component mean vector was estimated by three different "naive" approaches :
Complete Cases (CC), All Available Data (AAD), and Last Value (LV) and the results were compared to those obtained using a maximum likelihood approach (EM(a)).

Note here that CC, AAD and LV approaches will not be valid since the data are not Missing Completely at Random (MCAR).
The estimates obtained under the four approaches are shown in Table 4.2 and in Figures 4.1 & 4.2 and could be summarised in a few points (in each case the maximum likelihood approach will be used as a "yardstick" to compare with other results, since it is the approach with most theoretical grounding).

The results of applying EM(a) to the two groups are shown in Figure 4.3.

(i) The CC approach is generally producing estimates of the mean which are too low - this is what could be expected, since those cases reaching the twelve-week assessment are those whose blood pressure was fairly well controlled even by Week 4.
During the study, cases found to have excessively high

Table 4.2 : Comparison of CC, AAD, LV and EM(a) (Mean(s.e.))

Ketanserin Group

	Time Point				
	1	2	3	4	5
CC	105.8(0.6)	97.4(0.7)	93.8(0.7)	92.9(0.7)	92.0(0.8)
AAD	108.7(0.7)	101.5(0.8)	99.5(0.9)	95.3(0.9)	91.9(0.8)
LV	108.7(0.7)	102.0(0.8)	100.2(0.9)	99.9(1.0)	99.2(1.0)
EM(a)	108.7(0.7)	101.6(0.8)	99.8(0.9)	98.4(1.0)	96.3(0.9)

Metoprolol Group

	Time Point				
	1	2	3	4	5
CC	106.6(0.6)	95.3(0.8)	92.7(0.7)	92.1(0.8)	92.7(0.7)
AAD	108.0(0.6)	97.5(0.8)	96.3(0.8)	93.6(0.8)	93.1(0.7)
LV	108.0(0.6)	97.8(0.8)	96.6(0.8)	96.2(0.8)	96.6(0.8)
EM(a)	108.0(0.6)	97.5(0.8)	96.5(0.8)	95.1(0.8)	95.1(0.7)

Figure 4.1 : Comparison of CC. AAD and EM(a) (Ketanserin)

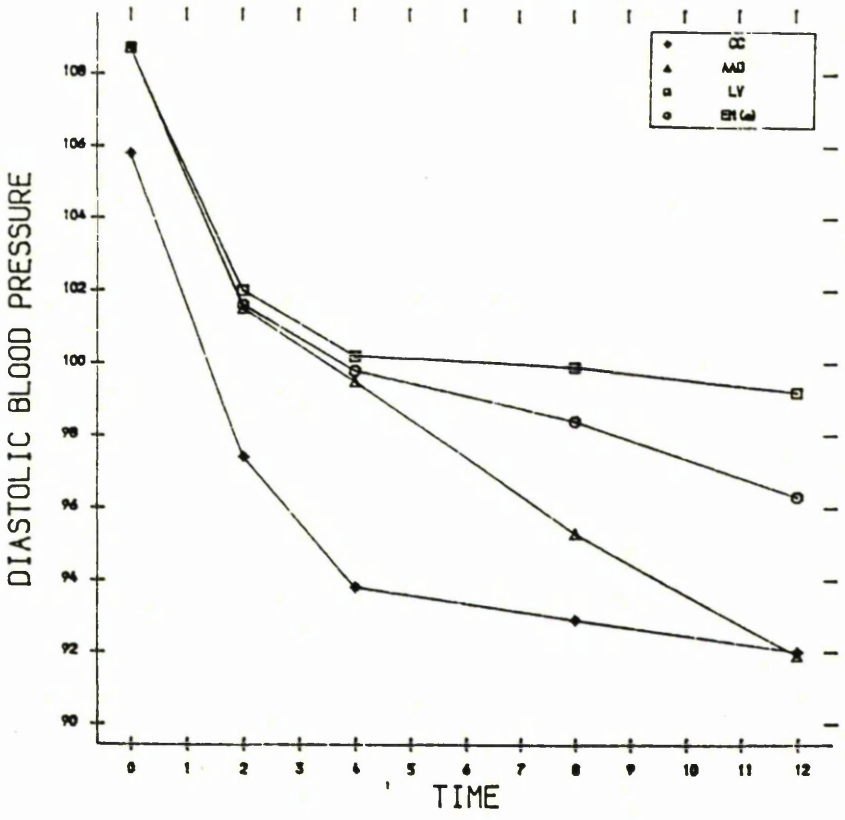


Figure 4.2 : Comparison of CC. AAD and EM(a) (Metoprolol)

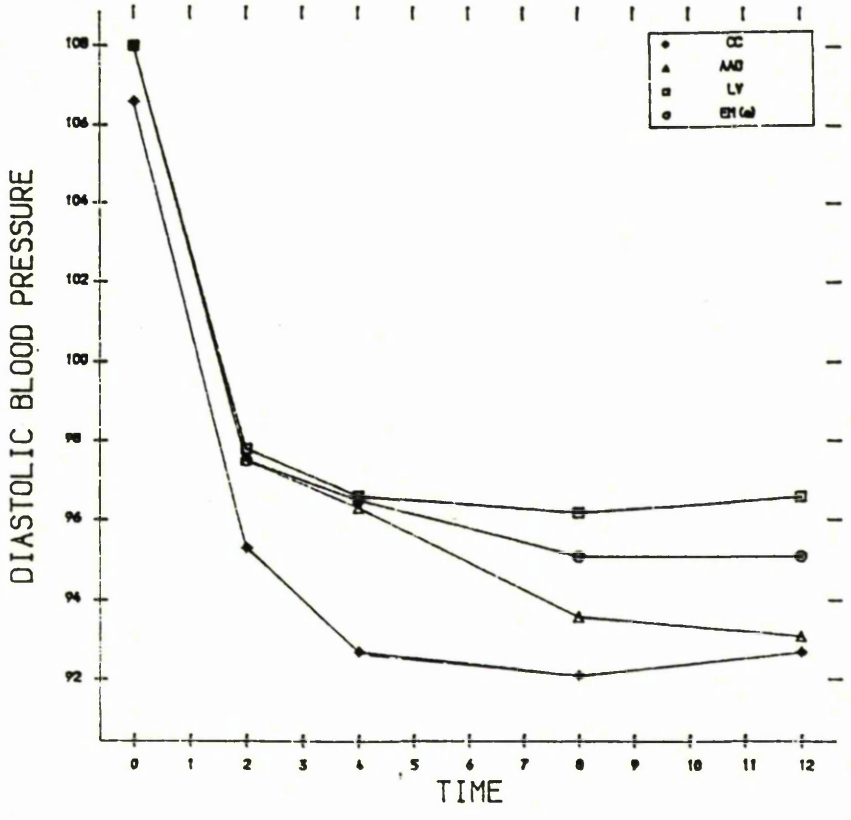
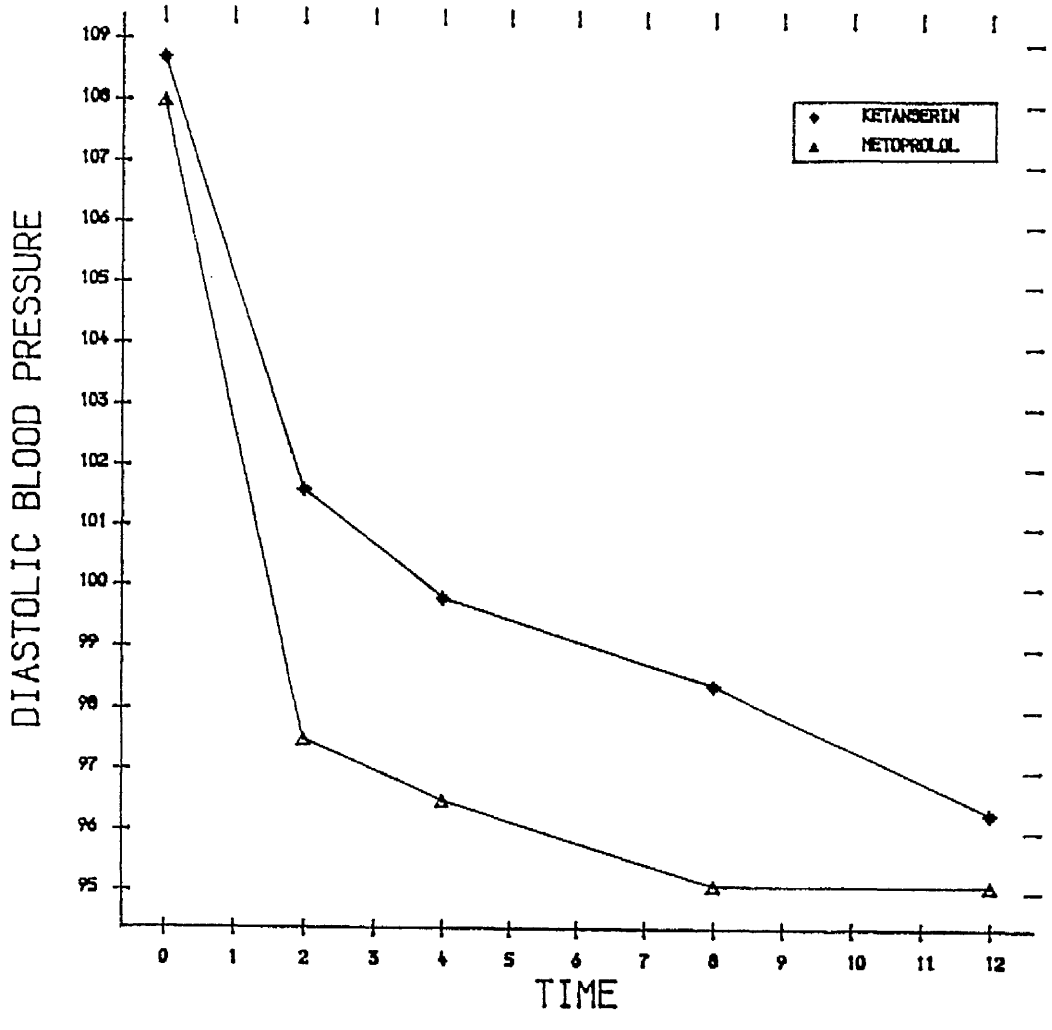


Figure 4.3 : Results of EM(a) on Ketanserin and Metoprolol



blood pressure were excluded from, at the very least, the final (Week 12) visit.

(ii) The AAD approach agreed with the EM(a) approach at Week 0, where the data was complete, and gave results close to the CC approach for the Week 12 data. This is natural. At Week 0, the data are complete, so that the sample mean of these data will be optimal in terms of maximising the likelihood. At Week 12, the cases with observed data tend to be those who have been present throughout the study (i.e. the complete cases), and so the CC and AAD approaches give similar results.

(iii) The LV approach, again, agreed with the EM(a) approach at the first time point for reasons as given above for AAD. However, as time progresses, the LV mean estimates diverge from the maximum likelihood values, overestimating the mean values.

Again, this is not surprising. By the nature of the Last Value approach, previous (generally larger) data values are being used to replace missing values, and the proportion of missing data increases through time.

(iv) From the maximum likelihood results, the reason for the larger proportion of withdrawals in the Ketanserin group can be seen. Although, by the end of the study, the average blood pressure reductions are approximately equal in the two groups (especially if corrected for the slight imbalance at randomisation), the time courses for the two drugs appear to be quite different.

For Ketanserin, there is a steady fall in blood pressure throughout the study, whereas, for Metoprolol, there is an initial rapid blood pressure reduction over the first two weeks, with little change seen over the remainder of the study. This would imply that at the four-week and eight-week points in the study, the Ketanserin group will have proportionately more cases with relatively high blood pressure than in the Metoprolol group, so leading to the compulsory removal of a greater number of cases, despite the fact that, eventually, the drug-effects level out.

Model Checking : To assess the suitability of the Multivariate Normal model, transformed distance plots (as discussed earlier) were produced for each group separately. These are shown in Figures 4.4 & 4.5. Recall that if a Multivariate Normal model was appropriate, then approximately one twentieth of the data points could be expected to lie outside the "envelopes" shown on the plots. (To illustrate the type of plots expected under a Multivariate Normal (MVN) model, several MVN data sets were simulated and deletion of observations was performed to give the same set of patterns as seen in the two groups. Transformed distance plots were then produced. One typical plot is shown in Figure 4.6).

From the transformed distance plots, it can be seen that the MVN model does not appear to be ideal. As a first stage, various transformations were used in an attempt to "Normalise" the data. The results of applying log-transformation and square-root transformation can be seen in Figures 4.7-4.10. The transformations appeared to have done little to improve the non-Normality of the data.

Due to the assessments made above, robust techniques were applied. For each group separately :

(i) MVt models were fitted for various degrees of freedom. The maximised log-likelihoods for each of the fitted models are shown in Table 4.3 and in Figures 4.11 & 4.12 . A suitable integer value of degrees of freedom was chosen to give the largest maximised log-likelihood (for Ketanserin, 5 degrees of freedom, and for Metoprolol 8 degrees of freedom).

(ii) CN models were fitted over a grid of values for δ and λ as defined earlier and the combination was, again, chosen to maximise the log-likelihood function. Much work was involved in this since no prior information was available about either of the parameters. The maximising combinations were found to be :

For Ketanserin : $\delta = 0.09$ $\lambda = 0.27$

For Metoprolol : $\delta = 0.23$ $\lambda = 0.22$

The maximum likelihood estimates for the means and standard deviations under the chosen MVt and CN models are shown in

Figure 4.4 : Transformed Distance Plot for Ketanserin

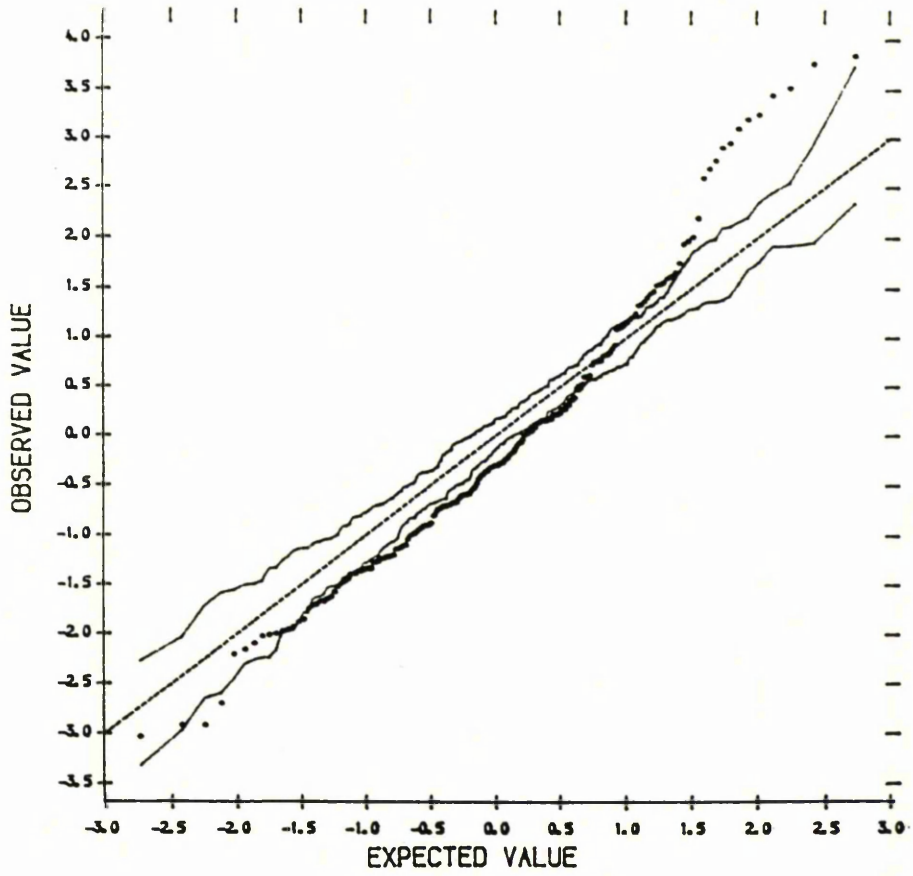


Figure 4.5 : Transformed Distance Plot for Metoprolol

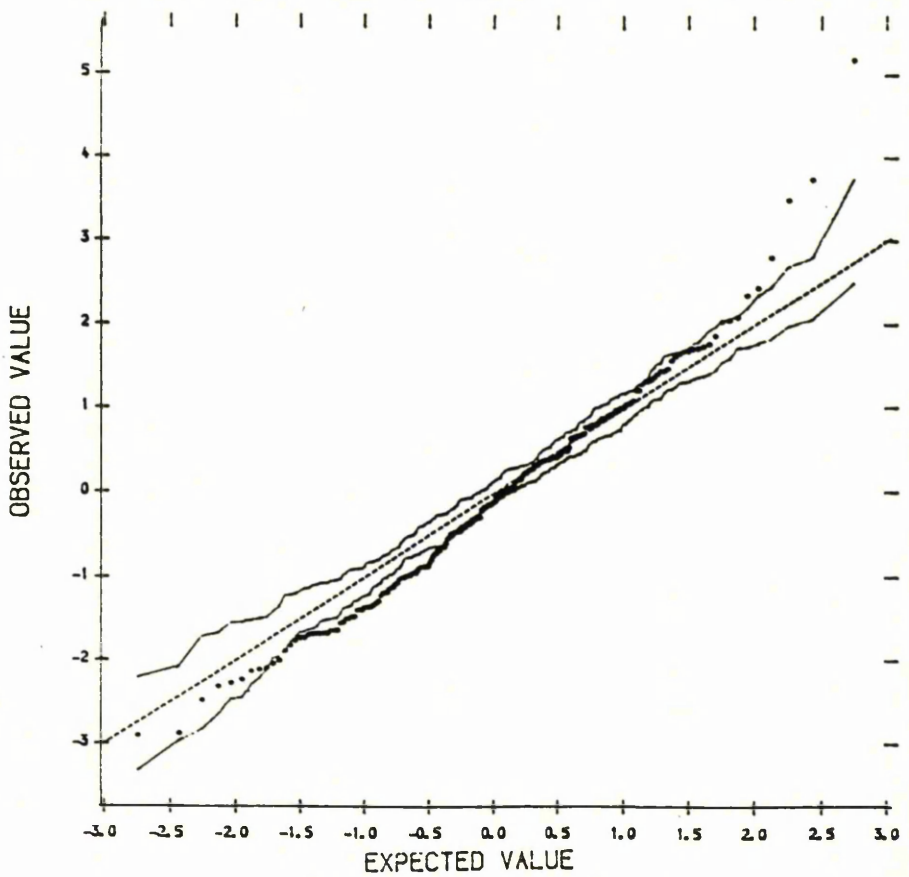


Figure 4.6 : Typical Transformed Distance Plot for MVN Data

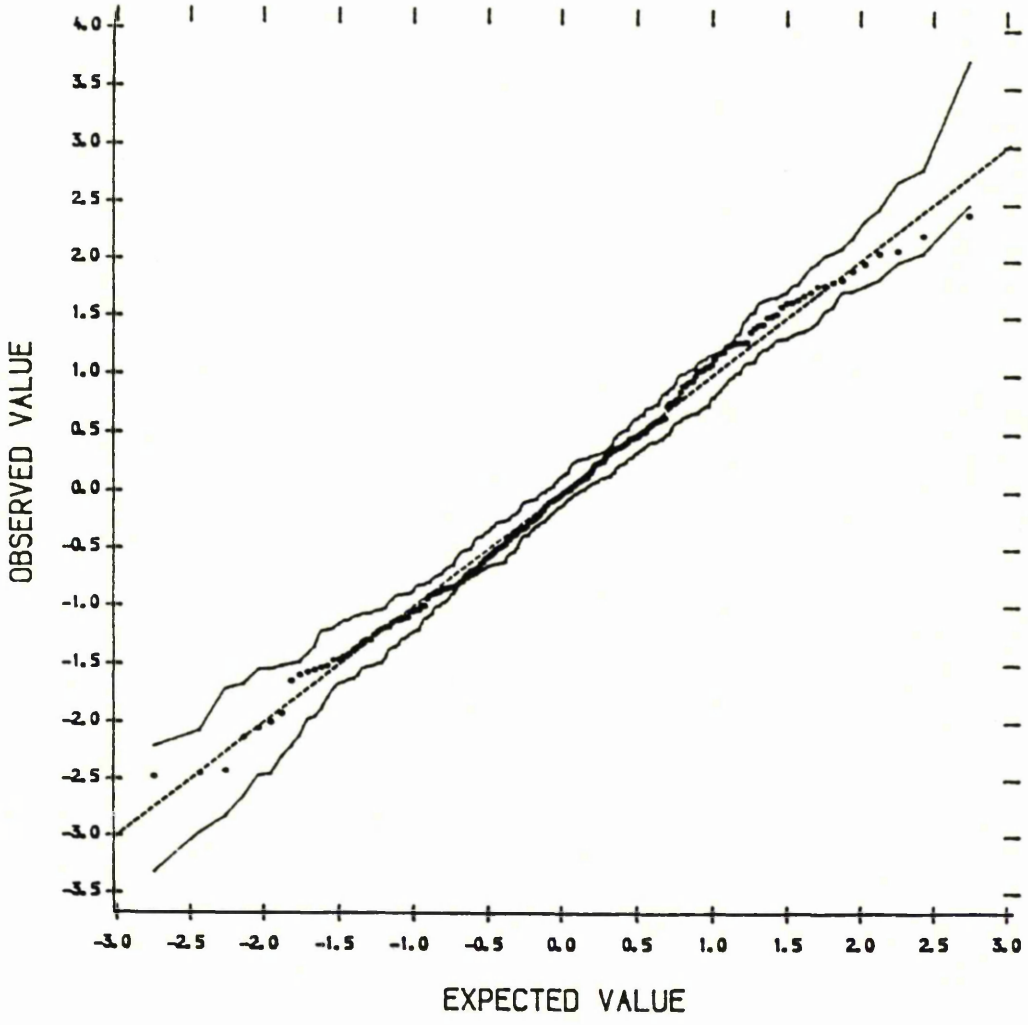


Figure 4.7 : T.D. Plot for $\log_e$(Ketanserin)

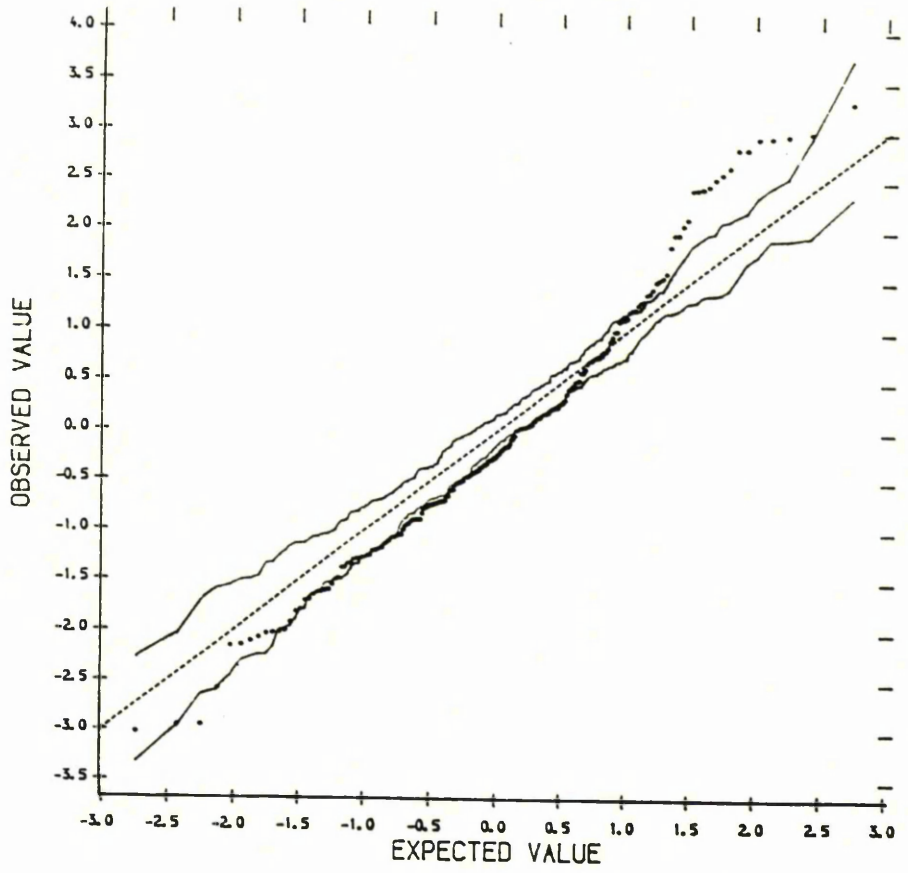


Figure 4.8 : T.D. Plot for $\log_e$(Metoprolol)

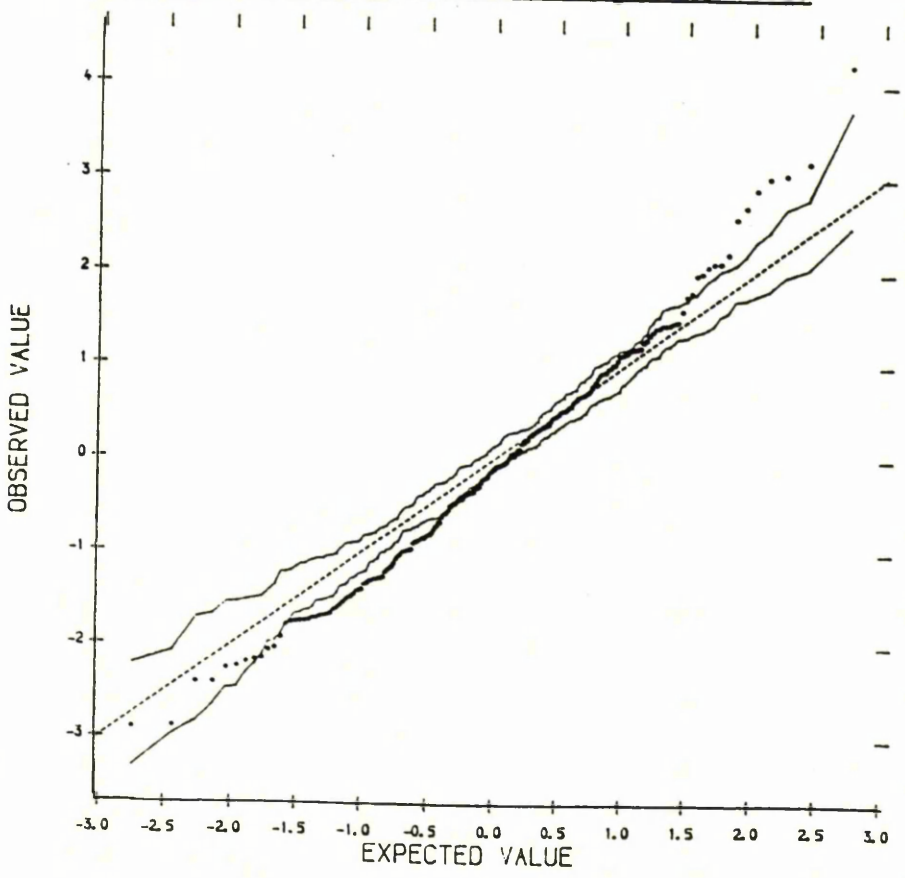


Figure 4.9 : T.D. Plot for ✓(Ketanserin)

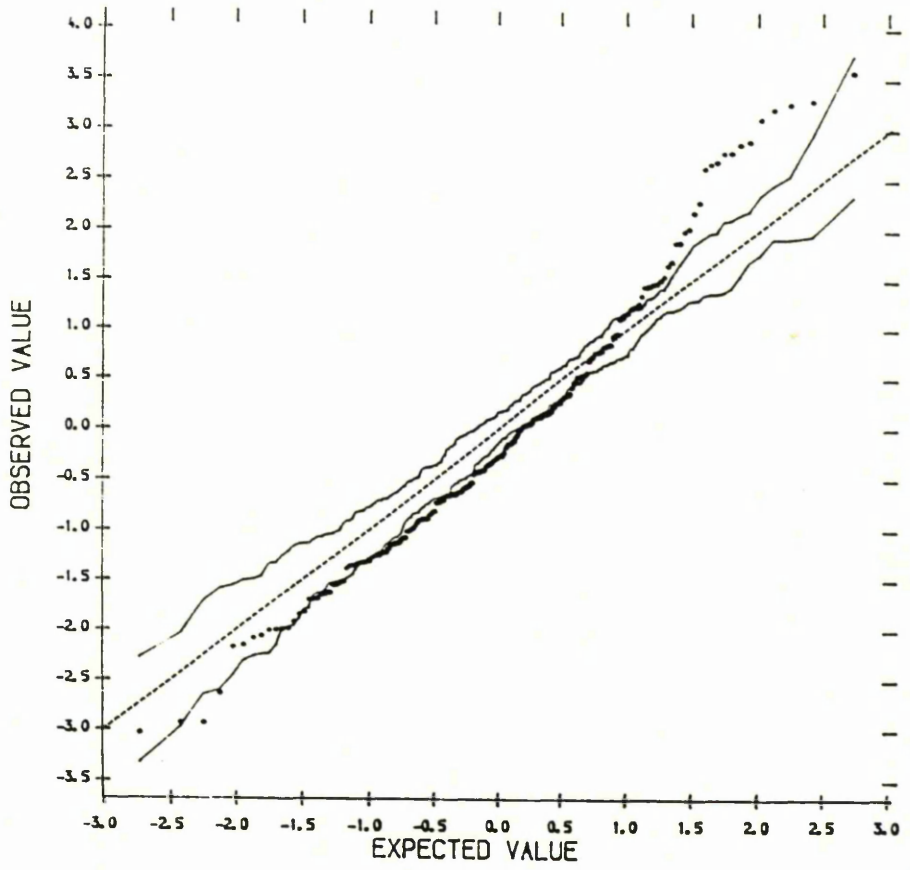


Figure 4.10 : T.D. Plot for ✓(Metoprolol)

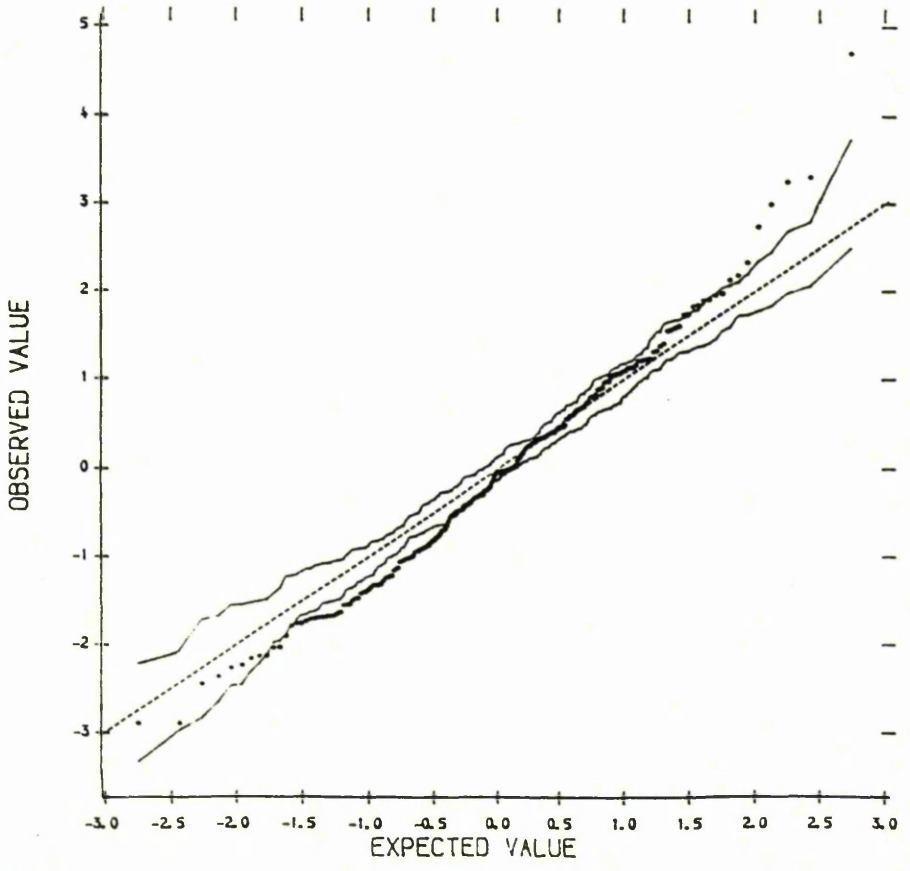


Table 4.3 : Fitting MVt Models - Maximised Log-likelihood as a Function of df

Degrees of Freedom	Ketanserin Group	Metoprolol Group
1	-3329.8105	-3593.4643
2	-3288.0618	-3546.0497
3	-3275.6994	-3529.8560
4	-3271.5633	-3522.7995
5	**-3270.5540**	-3519.4491
6	-3270.8991	-3517.8616
7	-3271.8597	-3517.1952
8	-3273.0927	-3517.0385
9	-3274.4285	-3517.1736
10	-3275.7801	-3517.4780
11	-3277.1026	-3517.8797
12	-3278.3725	-3518.3352
13	-3279.5802	-3518.8176
14	-3280.7218	-3519.3098
15	-3281.7973	-3519.8011
20	-3286.2838	-3522.0817
25	-3289.6117	-3523.9684
30	-3292.1468	-3525.5010
50	-3298.1150	-3529.4139
75	-3301.6205	-3531.9115
100	-3303.5212	-3533.3295

Figure 4.11 : Maximised Log-Likelihoods Produced by Fitting Various MVt Models to the Ketanserin Data

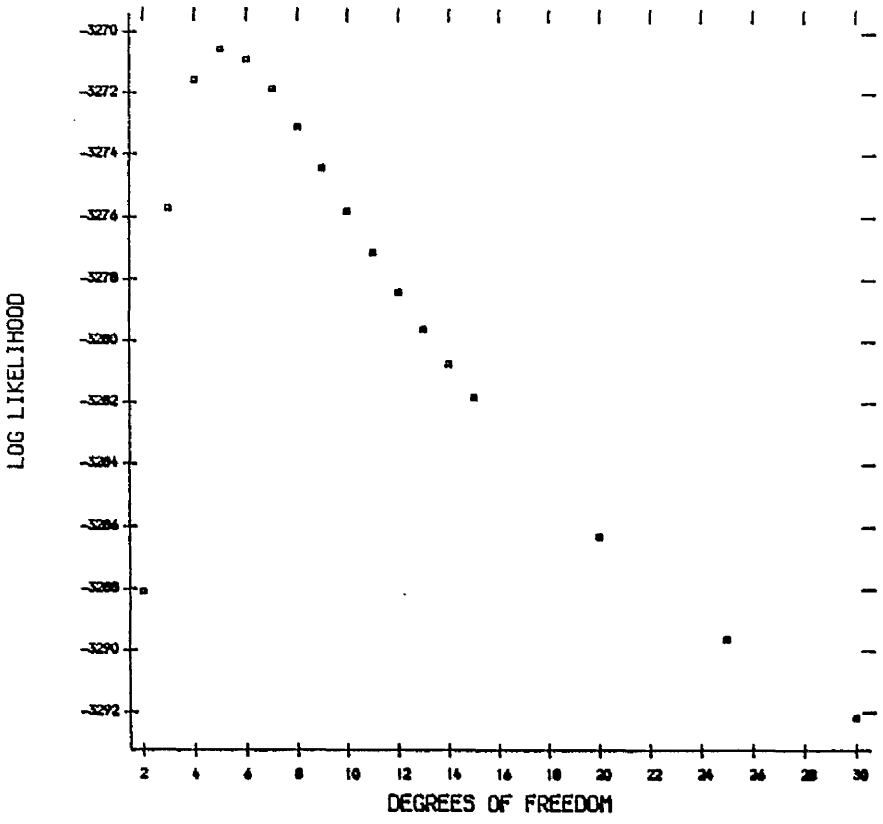


Figure 4.12 : Maximised Log-Likelihoods Produced by Fitting Various MVt Models to the Metoprolol Data

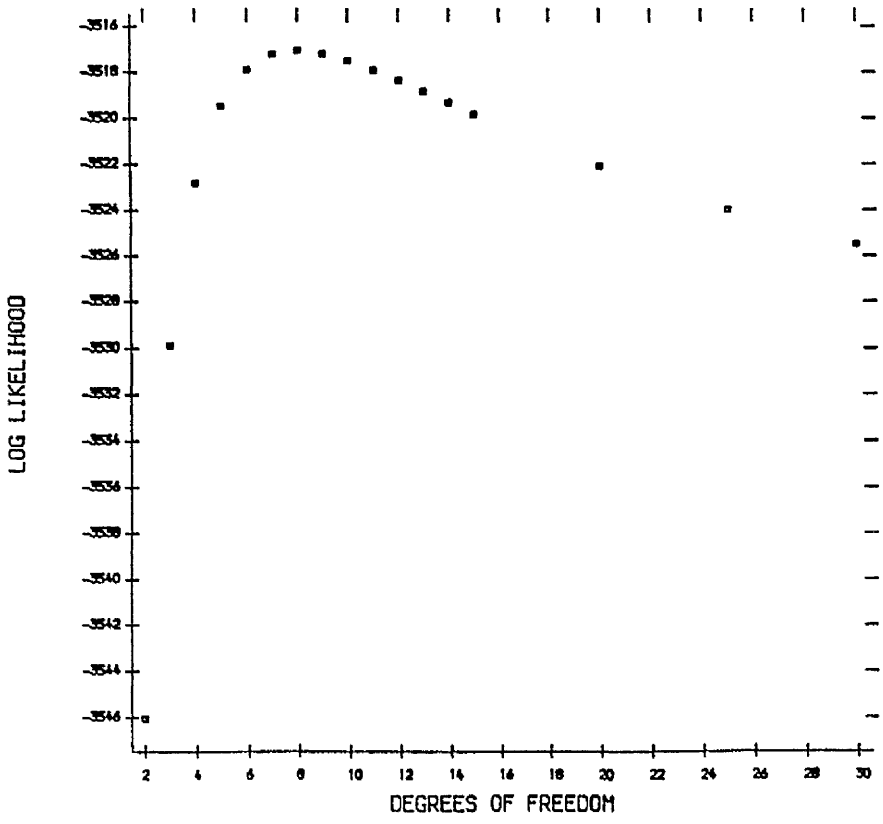


Table 4.4 together with the results obtained in fitting model EM(a). A comparison of the methods is shown in Figure 4.13 & 4.14

For information, plots of the case-weights are shown as a function of d_i^2 for the two robust models (see Figures 4.15-4.18)

It can be seen that under the three different models, the maximum likelihood estimates for the group means are similar. However, the standard deviations are quite different, being substantially smaller under the two robust approaches. (Note that the quoted standard deviations refer to the whole population, including outliers, not just to the "uncontaminated" portion of the population). The reason for this is that whenever any calculations are performed with the robust methods, the more extreme values are downweighted, so that they have less effect on the parameter estimates (e.g. of the standard deviations, means, etc). The result is that the standard deviations are smaller and the means are also slightly smaller, since extreme values tend to be values which are "too large" rather than "too small".

One important point to emphasise is that if the ultimate aim after application of the robust methods was to compare the group means, it would be desirable to fit similar models to these groups, in the sense that for the MVt model, it would be desirable to fit models to the groups which had the same degrees of freedom and common Σ, while for the CN model, it would desirable to fit models with the same values of δ and λ and common Σ.

Although it has not been done here, there would be no theoretical difficulties in extending algorithm EM(b) to cover such models.

Do the Robust Methods Provide a Better Fit to the Data ?

The question in the title was addressed by performing likelihood ratio tests on the maximised log-likelihoods under the various models.

Let l_0 = maximised log-likelihood under the simpler model being considered (i.e. MVN)

and let l_1 = maximised log-likelihood under the more complex model being considered.

Table 4.4 : Comparison of Robust Methods to EM(a)
(Mean(s.d.))

Ketanserin Group

	Time Point				
	1	2	3	4	5
EM(a)	108.7(10.1)	101.6(11.9)	99.8(13.8)	98.4(13.3)	96.3(12.1)
MVt	107.0(7.7)	99.6(9.5)	97.8(10.8)	96.2(10.9)	94.4(9.5)
CN	107.2(7.2)	99.8(9.0)	98.2(10.1)	96.9(10.5)	94.8(9.2)

Metoprolol Group

	Time Point				
	1	2	3	4	5
EM(a)	108.0(8.4)	97.5(11.1)	96.5(11.5)	95.1(11.5)	95.1(10.1)
MVt	107.3(7.3)	97.4(9.9)	96.2(10.0)	95.2(10.4)	94.7(9.3)
CN	107.5(7.5)	97.5(10.3)	96.4(10.4)	95.1(10.6)	94.9(9.6)

Figure 4.13 : Comparison of MVt, CN and EM(a) (Ketanserin)

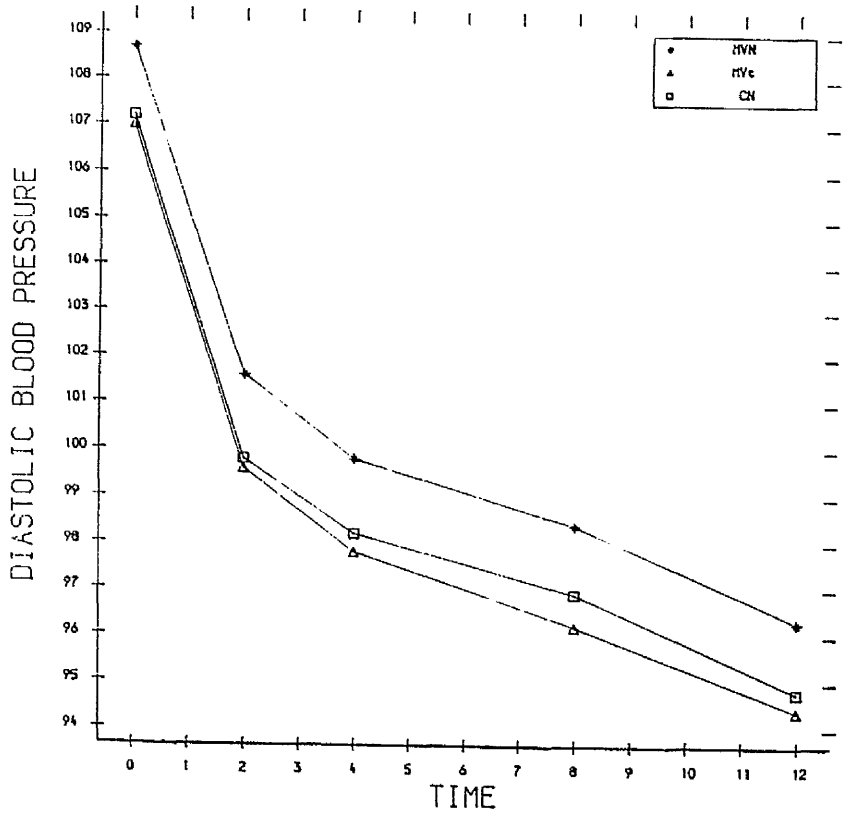


Figure 4.14 : Comparison of MVt, CN and EM(a) (Metoprolol)

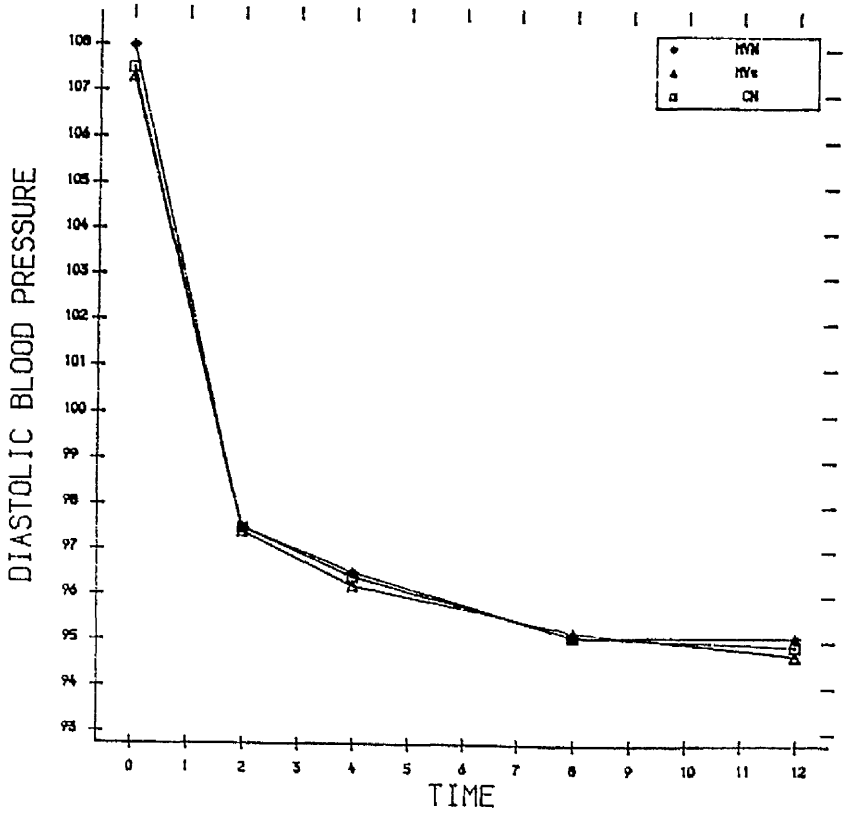


Figure 4.15 : w_i versus d_i^2. MVt Model (Ketanserin)

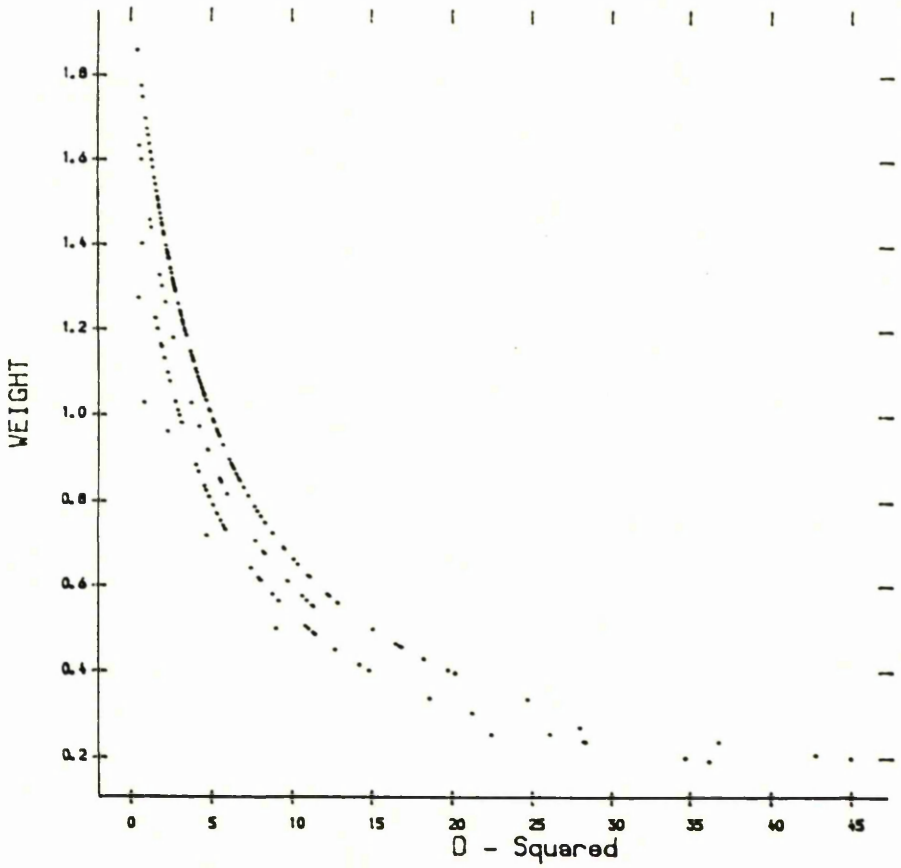


Figure 4.16 : w_i versus d_i^2. CN Model (Ketanserin)

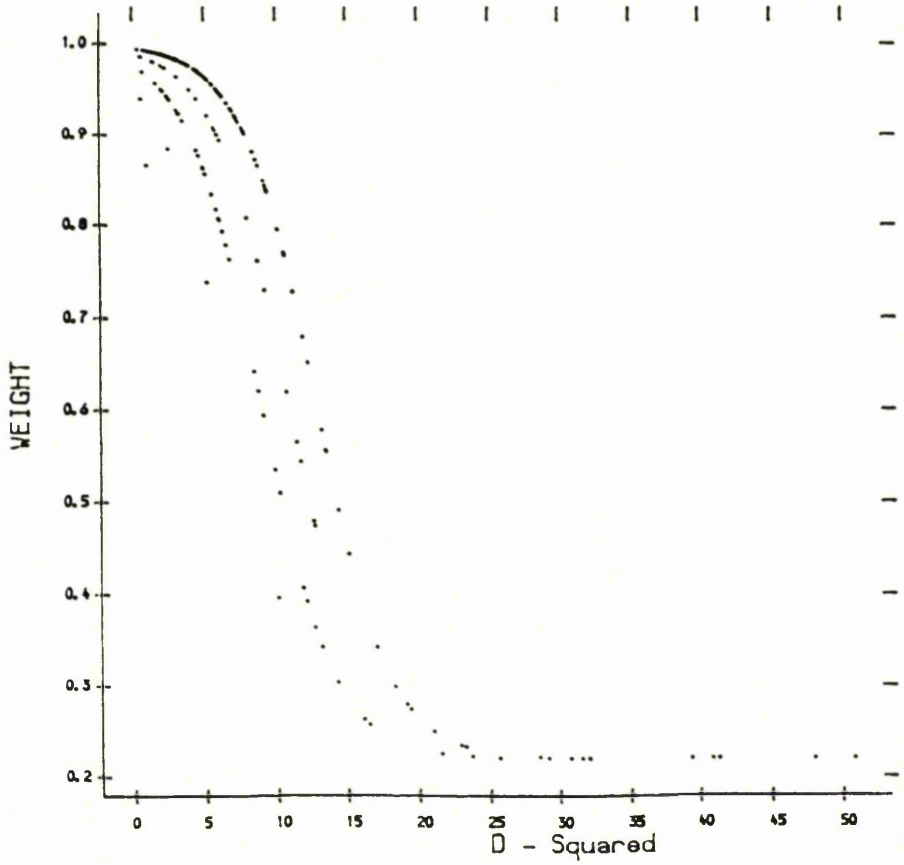


Figure 4.17 : w_i versus d_i^2. MVt Model (Metoprolol)

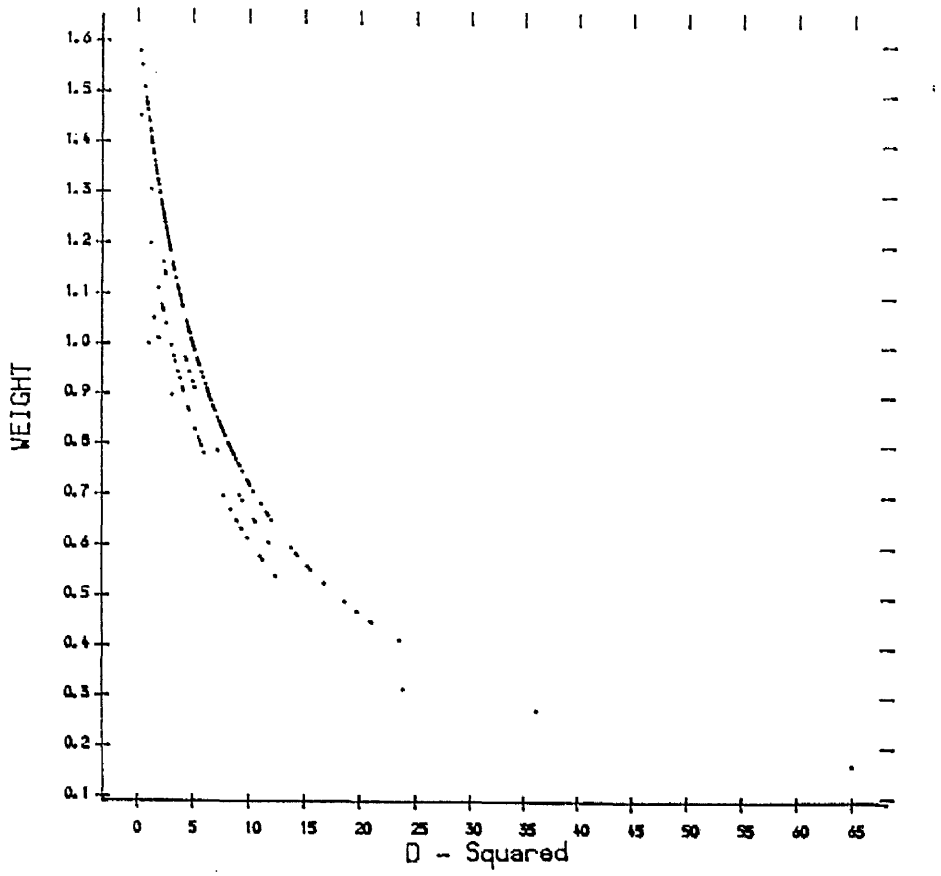
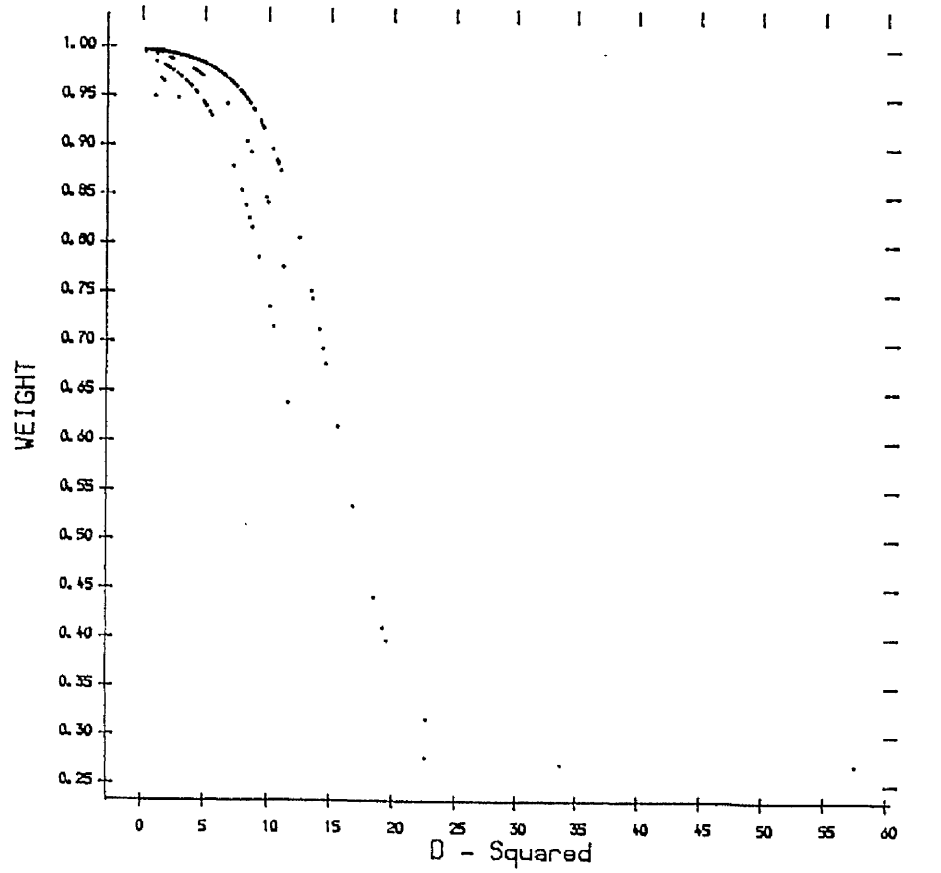


Figure 4.18 : w_i versus d_i^2. CN Model (Metoprolol)



The Ketanserin Group : the maximised log-likelihoods were as shown below.

Model	Maximised log-likelihood
MVN	-3309.881
MVt	-3270.554
CN	-3268.750

(a) Comparison of the MVN and MVt Models

$$2 \log(\lambda) = 2(l_1 - l_0) = 78.65$$

This test statistic is compared with $\chi^2(1 ; 0.95) = 3.841$.

Conclusion : the MVt model fits the data significantly better than the MVN model.

(b) Comparison of the MVN and CN Models

$$2(l_1 - l_0) = 82.26$$

Conclusion : the CN model appears to fit the data better than the MVN model.

It can be seen that there is little difference between the maximised log-likelihoods under the two robust models, so that probably the model with fewer parameters to fit, i.e the MVt Model, would be preferred.

The Metoprolol Group : the maximised log-likelihoods were as shown below.

Model	Maximised log-likelihood
MVN	-3538.415
MVt	-3517.039
CN	-3521.648

(a) Comparison of the MVN and MVt Models

$2 \log(\lambda) = 2(l_1 - l_0) = 42.75$

Compare this test statistic to $\chi^2(1 ; 0.95) = 3.841$.

Conclusion : The MVt model fits the data significantly better than the MVN model.

(b) Comparison of the MVN and CN Models

$2(l_1 - l_0) = 33.47$

Conclusion : The CN model appears to fit the data better than the MVN model.

Here it can be seen that the MVt model, despite fitting fewer parameters than the CN model, provides a better fit to the data.

The following section could be considered to be slightly controversial in the light of the results of the last section in that it will go back to fitting MVN models to the data. From the transformed distance plots one could argue that although the z_i fall outside the expected range in several cases, the only serious deviations are to be seen in the distribution tails, so that the results of fitting MVN models should not be too misleading. Also, since the sample sizes are large, the goodness of fit tests will be be sensitive even to small (and unimportant) departures from Normality.

The Comparison of Ketanserin and Metoprolol

Recall that earlier it was suggested that the maximised log-likelihoods obtained by implementation of the algorithms EM(a) and EM(b) could be used in order to test sequences of hypotheses of form

Stage 1 $H_0 : \mu_1, \mu_2, \Sigma$ vs $H_1 : \mu_1, \mu_2, \Sigma_1, \Sigma_2$,

followed up by

Stage 2 $H_0 : \mu , \Sigma$ vs $H_1 : \mu_1, \mu_2, \Sigma$

if the first null hypothesis could not be rejected.

The maximised log-likelihoods for the various models shown above for the Ketanserin vs Metoprolol problem are shown below.

Model	Maximised log-likelihood
$\mu_1, \mu_2, \Sigma_1, \Sigma_2$	-6848.295
μ_1, μ_2, Σ	-6858.937
μ, Σ	-6868.813

The Tests.

Stage 1 : $H_0 : \mu_1, \mu_2, \Sigma$ $H_1 : \mu_1, \mu_2, \Sigma_1, \Sigma_2$

$2 \log(\lambda) = 2(l_1 - l_0) = 21.28$

This is compared to $\chi^2(\frac{1}{2} p(p+1) ; 0.95)$

$= \chi^2(15 ; 0.95) = 25.00$.

Conclusion : There is no strong evidence that the covariance matrices are unequal. Therefore, proceed to Stage 2.

Stage 2 : $H_0 : \mu, \Sigma$ $H_1 : \mu_1, \mu_2, \Sigma$

$2 \log(\lambda) = 2(l_1 - l_0) = 19.75$

This is compared to $\chi^2(p ; 0.95) = \chi^2(5 ; 0.95) = 11.07$

Conclusion : There is some evidence that the group means are not equal.

Proceeding a stage further, how correct were the impressions that, despite the differing time - courses, the drugs produced a similar ultimate effect on the blood pressure ?

In an attempt to answer the above question, an approximate 95% confidence interval was produced for the difference between the overall changes in blood pressure (in mm Hg) in the two groups, i.e for

$c^T(\mu_1 - \mu_2)$ where $c^T = (1\ 0\ 0\ 0\ -1)$.

The Calculations

$c^T(\hat{\mu}_1 - \hat{\mu}_2) = 0.046$ mm Hg

$c^T(\ \mathrm{cov}(\hat{\mu}_1)\ )c = 0.6200$

$c^T(\ \mathrm{cov}(\hat{\mu}_2)\ )c = 0.5494$

Therefore the required interval is given by

$0.046 \pm N(0,1 ; 0.975) \sqrt{(0.6200 + 0.5494)}$ mm Hg

$= 0.046 \pm 2.120$ mm Hg

$= [-2.1 , 2.2]$ mm Hg

Conclusion : The conclusions from using maximum likelihood techniques to analyse this study would be that despite the time courses of the two drugs being different, there is no strong evidence that their ultimate effects are different. Even at the extremes of the interval above, there would be no difference of clinical relevance between the two treatments.

If, alternatively, one of the naive approaches had been employed, the results obtained could have been quite different. Looking at the differences between the average blood pressure reductions in the two groups, the answers obtained ranged from a difference of approximately 2 mm Hg in favour of ketanserin using the AAD approach, to a similar difference in favour of metoprolol using the LV approach, where each of these differences had approximately unit standard error. Thus different conclusions could have been obtained, dependent on the method of analysis employed. The reasons why the results from the maximum likelihood approach could be most readily accepted can be seen from the context of the problem. The very design of the study in question dictates that a large proportion of the missing data should be missing at random. An analysis which builds in that information will be preferable to one which does not.

Note : The estimates of means and standard errors under EM(b) are not given, as these are, for all practical purposes, the same are those obtained under EM(a).

4.2 : Rat Study (taken from Crepeau et al(1985)

Background : The data considered are those from a study of the effect of various doses of halothane on the responses of rats to induced irreversible myocardial ischaemia and subsequent infarction (leading to death in a large number of cases).

The halothane was administered in different doses to five groups of rats, and the resultant blood pressures were recorded through time (nine time points in all). The five groups were defined as :

Group 1	:	Controls	(11 cases)
Group 2	:	0.25% halothane	(10 cases)
Group 3	:	0.50% halothane	(11 cases)
Group 4	:	1.00% halothane	(11 cases)

Group 5 : 2.00% halothane (11 cases)

The Data : From the design of the study, the data were almost bound to be incomplete (clearly no data will be available after death !). The authors in the original paper carried out analyses based on assuming a "missing at random" set up, but did not state this assumption explicitly. Instead, they justified their analyses on the grounds that the blood pressure did not appear to be related to mortality, and that even the pharmacologist responsible for the experiment was unable to predict which rats would die by looking at their blood pressure responses. It is unclear from such statements whether a MCAR or a MAR assumption is implied.

(Note that for this example, since the data form a nested pattern, the use of iterative procedures would not be essential).

Following the approach of the original authors, the 2.00% halothane will be excluded from analyses, since no cases were completely observed in that group.

The patterns of missing data present in the remaining groups are shown in Table 4.5.

Due to the small number of cases per group, it would not be possible to perform the full likelihood ratio testing procedure described earlier, since it would not be possible to fit the full μ_i , Σ_i model (i.e. cannot apply EM(a) to each group separately).

Comparison of the Results of EM(b) and Some "Naive" Approaches

The means and standard errors obtained from the CC, AAD and EM(b) approaches are shown in Table 4.6 , with plots shown in Figures 4.19-4.22. The results could be summarised as follows:

As in the previous example, approaches EM(b) and AAD agreed at the first time point, while CC and AAD agreed at the final time point. Recall that the reasons were

(i) The first time point included data from all individuals, and so the average of these (as in AAD) would produce unbiased estimates of the four group means, and so would agree with the maximum-likelihood approach (EM(b)).

(ii) At the last time point, the cases still available would be those present throughout, and so the CC and AAD approaches would produce the same results.

Table 4.5 : Missing Data Patterns

Control Group

Minutes									Number of Cases
1	5	10	15	30	60	120	180	240	
1	1	1	1	1	1	1	1	1	6
1	1	1	1	1	1	1	0	0	1
1	1	0	0	0	0	0	0	0	4
									11

0.25% Halothane Group

Minutes									Number of Cases
1	5	10	15	30	60	120	180	240	
1	1	1	1	1	1	1	1	1	6
1	1	1	1	1	1	1	0	0	1
1	1	1	1	1	1	0	0	0	1
1	1	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	1
									10

0.50% Halothane Group

Minutes									Number of Cases
1	5	10	15	30	60	120	180	240	
1	1	1	1	1	1	1	1	1	5
1	1	1	1	1	1	1	0	0	1
1	1	1	1	1	1	0	0	0	1
1	1	1	1	0	0	0	0	0	1
1	1	0	0	0	0	0	0	0	3
									11

1.00% Halothane Group

Minutes									Number of Cases
1	5	10	15	30	60	120	180	240	
1	1	1	1	1	1	1	1	1	8
1	1	1	1	1	1	0	0	0	1
1	1	0	0	0	0	0	0	0	2
									11

Table 4.6 : Comparison of CC, AAD and EM(b) (Mean(s.e.))

Group 1 : Control

	Time Point				
	1	2	3	4	5
	6	7	8	9	
CC	100.8(3.8)	101.3(7.9)	99.2(5.6)	96.3(4.0)	99.2(5.1)
	107.9(2.9)	103.8(3.1)	101.7(2.1)	100.5(4.0)	
AAD	101.8(7.0)	99.4(8.0)	101.4(5.2)	98.6(4.1)	100.7(4.6)
	108.2(2.5)	107.1(4.3)	101.7(2.1)	100.5(4.0)	
EM(b)	101.8(5.7)	99.4(8.1)	98.6(8.0)	95.3(7.8)	97.7(5.8)
	105.5(5.4)	104.0(5.7)	101.5(5.8)	100.3(5.9)	

Group 2 : 0.25% Halothane

	Time Point				
	1	2	3	4	5
	6	7	8	9	
CC	103.8(5.7)	103.0(10.1)	91.7(8.8)	90.4(9.2)	108.8(5.4)
	104.2(5.0)	107.9(5.5)	105.4(4.7)	108.8(5.3)	
AAD	103.8(4.0)	98.1(7.6)	99.1(9.2)	93.4(7.2)	107.2(4.1)
	104.1(4.0)	107.1(4.7)	105.4(4.7)	108.8(5.3)	
EM(b)	103.8(6.0)	100.3(8.6)	98.5(7.8)	93.1(7.7)	106.2(5.9)
	103.4(5.5)	107.2(5.9)	104.5(6.0)	108.0(6.0)	

Group 3 : 0.50% Halothane

	Time Point				
	1	2	3	4	5
	6	7	8	9	
CC	94.0(4.6)	95.5(5.3)	93.0(5.3)	93.1(7.2)	93.5(7.8)
	98.0(6.9)	96.2(6.8)	94.0(7.2)	90.5(6.5)	
AAD	89.1(4.4)	83.7(8.0)	89.3(8.2)	89.1(9.1)	95.7(5.6)
	99.6(4.9)	99.8(6.6)	97.1(6.7)	90.5(6.5)	
EM(b)	89.1(5.7)	83.7(8.1)	85.2(7.6)	85.0(7.5)	86.6(5.7)
	91.4(5.3)	91.8(5.8)	89.7(5.8)	86.0(6.0)	

Group 4 : 1.00% Halothane

	Time Point				
	1	2	3	4	5
	6	7	8	9	
CC	81.6(9.0)	81.6(12.8)	87.5(9.5)	86.2(10.2)	86.9(8.3)
	85.7(8.2)	79.4(8.6)	75.3(9.2)	75.0(8.5)	
AAD	82.0(7.6)	79.3(10.5)	88.9(8.5)	87.8(9.1)	88.1(7.4)
	86.4(7.3)	79.4(8.6)	75.3(9.2)	75.0(8.5)	
EM(b)	82.0(5.7)	79.3(8.1)	86.7(7.4)	85.4(7.2)	85.5(5.5)
	84.3(5.1)	78.5(5.6)	74.7(5.6)	74.2(5.5)	

Figure 4.19 : Comparison of CC, AAD and EM(b) (Control)

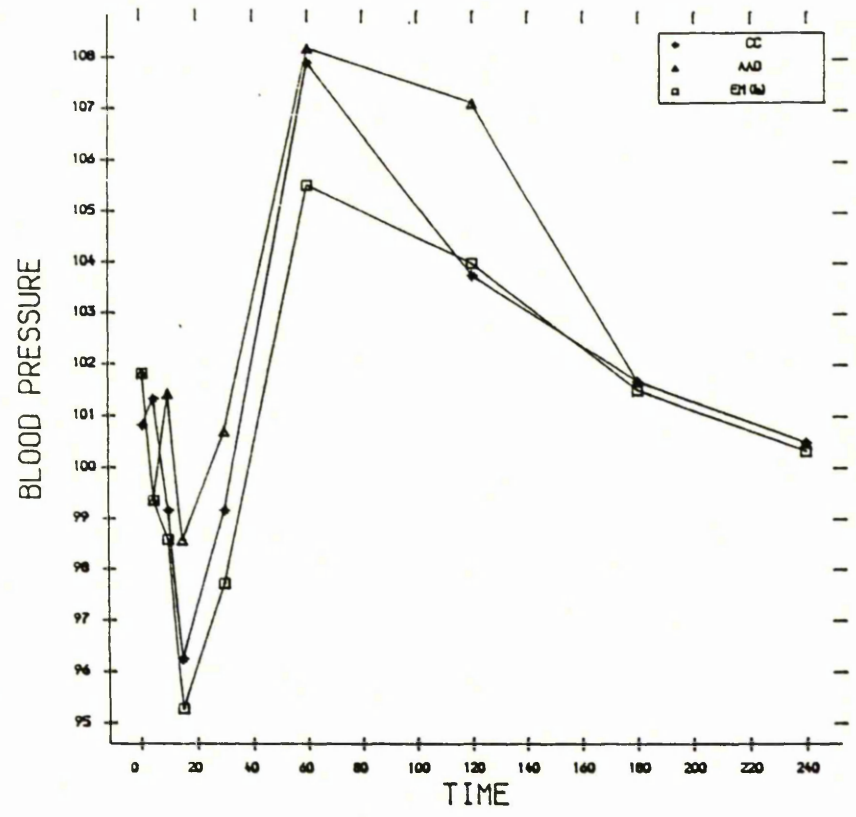


Figure 4.20 : Comparison of CC, AAD and EM(b) (0.25% Group)

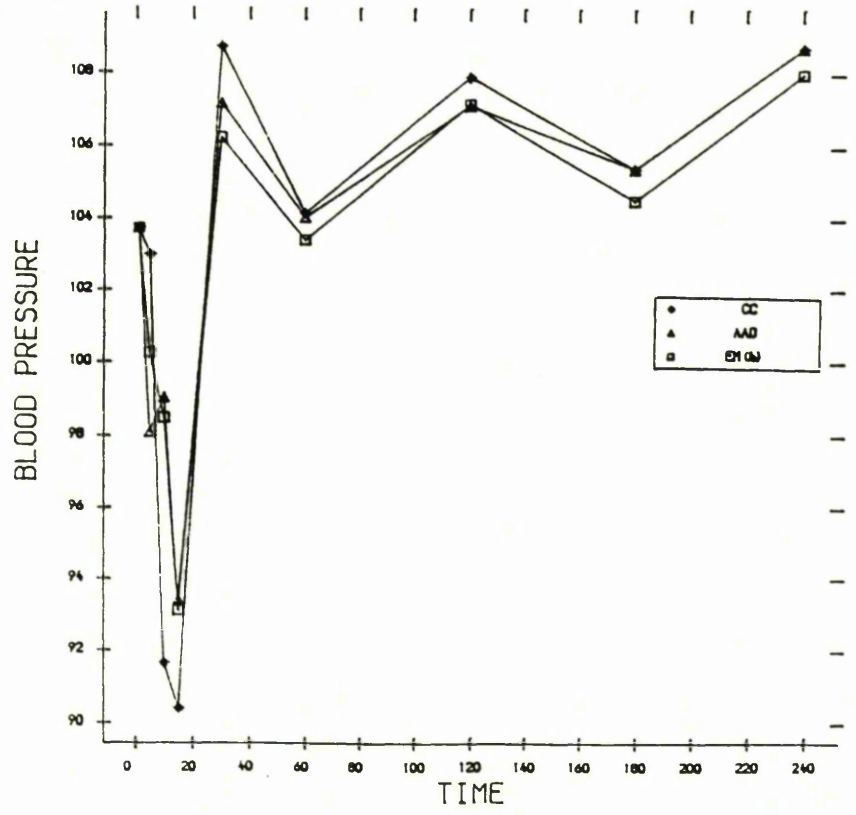


Figure 4.21 : Comparison of CC. AAD and EM(b) (0.50% Group)

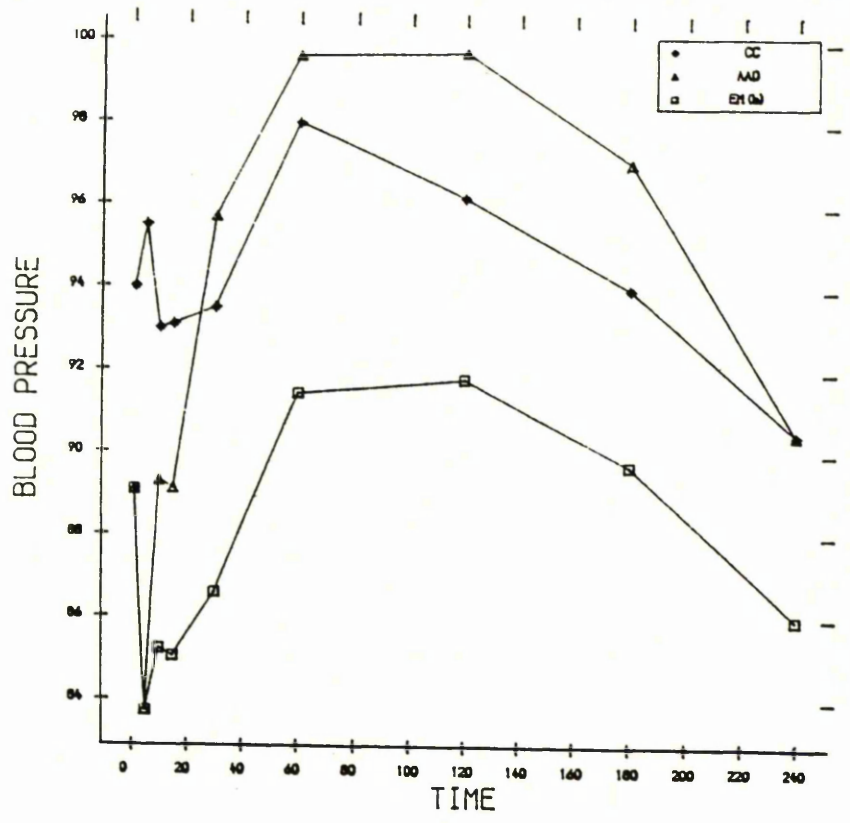
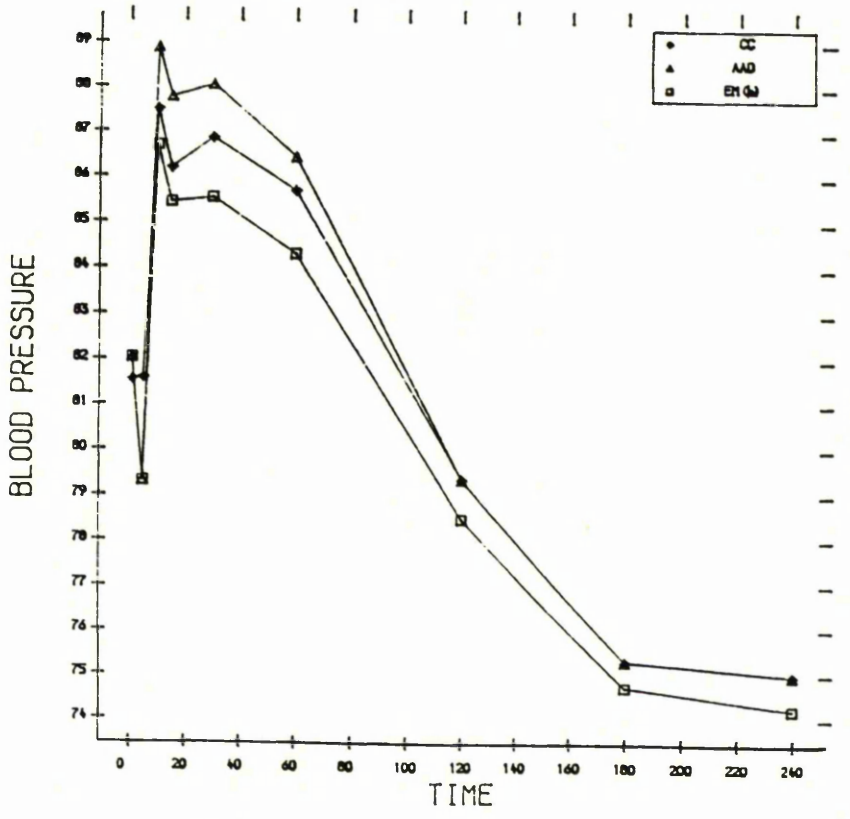


Figure 4.22 : Comparison of CC. AAD and EM(b) (1.00% Group)



The estimated means obtained from using the CC and AAD approaches tended to be higher than those obtained from using approach EM(b). This could be a product of some mechanism whereby individuals with lower blood pressure were being removed. Alternatively, these could merely be chance findings due to the general lack of data (any observed differences between the means calculated under the different approaches could be explained away in terms of the large calculated standard errors).

The results of EM(b) are compared for the four groups in Figure 4.23.

Model Assessment : As mentioned earlier, it was not possible to fit separate models to the four groups. As a result, it was necessary to produce a transformed distance plot (Figure 4.24) using the weights obtained in the procedure EM(b). From the plot, there appears to be no strong evidence that a MVN model is inappropriate.

Comparison of the Groups

The maximised log-likelihoods obtained under the various models are shown below.

Model	Maximised log-likelihood
μ_i , Σ_i	Cannot be fitted (lack of data)
μ_i , Σ	-1063.022
μ , Σ	-1082.097

The Test

$$2\log(\lambda) = 2(l_1 - l_0) = 46.15$$

Refer this to a $\chi^2(27)$ distribution, where $\chi^2(27 ; 0.95) = 40.11$.

Conclusion : There is some evidence that not all four group means are equal, i.e. there is evidence that the different doses of halothane produced different effects on the rats' blood pressures.

On examining the results for the four groups, it would appear that much of the difference between their means could be

Figure 4.23 : Results of EM(b) on All Groups

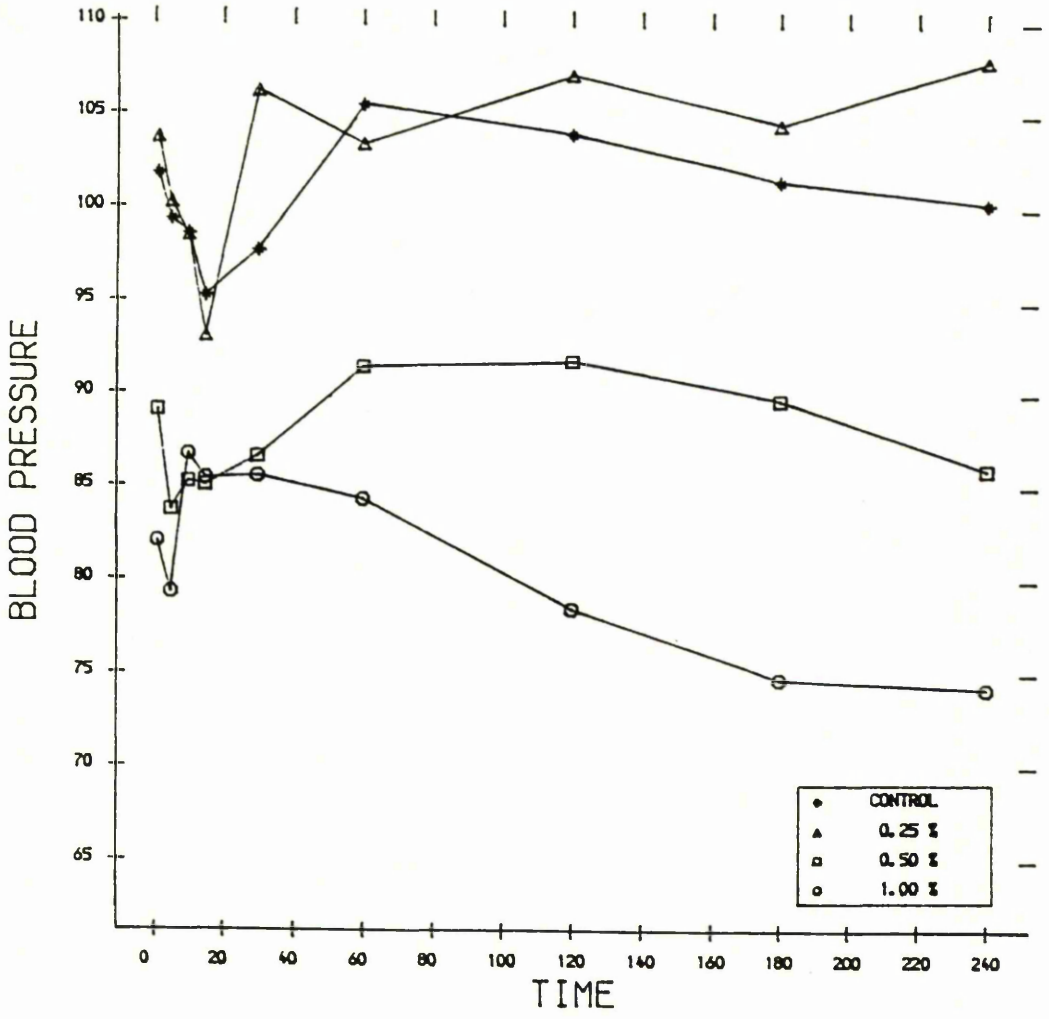
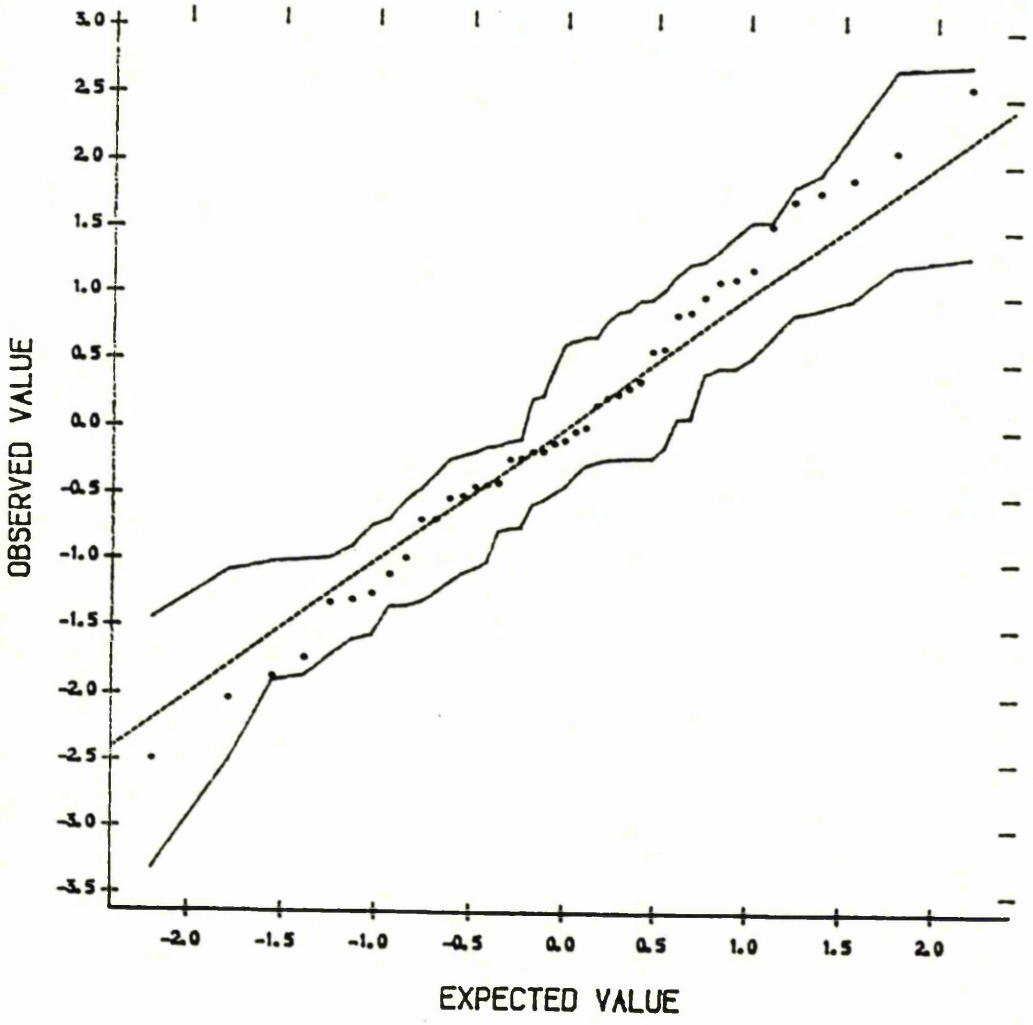


Figure 4.24 : T.D. Plot for All Groups (from EM(b))



explained by the differences in the initial blood pressures, in the sense that if linear adjustment was made for the differences in the initial pressures, there would be little difference between the blood pressure curves.

To investigate this, individuals had their initial blood pressure measurement subtracted from each of their subsequent measurements. The analyses described earlier were then repeated using these "adjusted" blood pressures for time points two onwards. This resulted in the loss of the one case for the 0.25% Halothane group who only had data at the first time point. The results obtained were as shown below.

Model	Maximised log-likelihood
μ_i , Σ_i	Cannot be fitted (lack of data)
μ_i , Σ	-893.743
μ , Σ	-905.743

The Test

$$2\log(\lambda) = 2(l_1 - l_0) = 24.00$$

Refer this to a $\chi^2(24)$ distribution, where $\chi^2(24 ; 0.95) = 36.42$

Conclusion : It would appear that after adjustment for the differing starting values, there are no significant differences among the blood pressure curves for the different groups. The estimated means under model EM(b) are shown in Table 4.7 and plotted in Figure 4.25.

It should be noted that the conclusions reached for the raw data differ from those in the original paper. There, the authors concluded, on the basis of a Score test, that there were no significant differences among the group means. This discrepancy in results probably reflects, more than anything else, the general lack of data. Both the Likelihood Ratio test and the Score test rely on approximate chi-squared distributions for their test statistics. It is likely that in neither case would these distributions be particularly appropriate, since there was so little data available. The most reasonable conclusion overall would probably, unfortunately, be one of uncertainty as to whether the groups were different.

Table 4.7 : Rat Study , Adjusted Data , EM(b) (Mean(s.e.))

Group 1 : Control

	Time Point				
	1	2	3	4	5
	6	7	8	9	
EM(b)	0.0(0.0)	-2.5(4.5)	-4.6(7.5)	-7.5(6.5)	-5.5(5.1)
	-2.2(5.0)	1.2(4.0)	-1.3(3.9)	-2.5(5.0)	

Group 2 : 0.25% Halothane

	Time Point				
	1	2	3	4	5
	6	7	8	9	
EM(b)	0.0(0.0)	-3.8(5.0)	-4.5(7.1)	-10.2(6.2)	3.2(5.0)
	0.5(4.8)	4.2(4.0)	1.4(3.8)	5.1(4.9)	

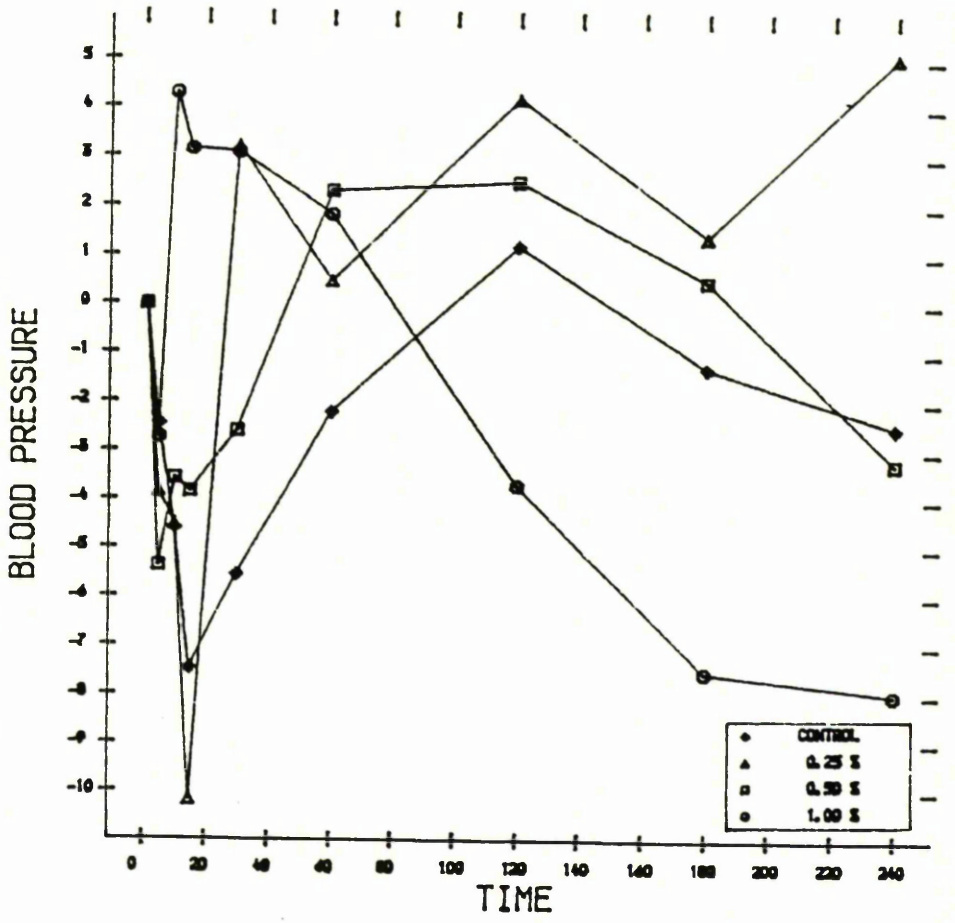
Group 3 : 0.50% Halothane

	Time Point				
	1	2	3	4	5
	6	7	8	9	
EM(b)	0.0(0.0)	-5.3(4.5)	-3.6(7.0)	-3.8(6.1)	-2.6(5.0)
	2.3(4.8)	2.5(4.1)	0.5(3.9)	-3.2(5.1)	

Group 4 : 1.00% Halothane

	Time Point				
	1	2	3	4	5
	6	7	8	9	
EM(b)	0.0(0.0)	-2.7(4.5)	4.3(6.6)	3.1(5.8)	3.1(4.6)
	1.8(4.5)	-3.7(3.7)	-7.5(3.5)	-8.0(4.5)	

Figure 4.25 : Results for EM(b) on the Adjusted Data



4.3 : The "Third Drug" Study (see Herrick et al(1989))

Background : a common mode of treatment for individuals with hypertension is by the use of so-called "Beta-blockers" and "Diuretics", often in combination. However, commonly, a proportion of cases will not have their blood pressure sufficiently controlled by such treatment, and will have some "third drug" added in.

The aim of this study was to compare four possible agents which could be added in at this stage, namely Placebo, Captopril, Hydralazine and Nifedipine. (The beta-blocker and the diuretic were, respectively, atenolol and bendrofluazide.)

Patients to take part in the study, after an initial run-in period on only the two baseline treatments, were randomised to receive one of the four possible "third drugs". Their blood pressure was then monitored through increasing dose levels of their third drug, until their blood pressure was controlled (or until a preset maximum dose was reached), from which point the patient remained under study for a further six weeks. The endpoint of the study was with the assessment of blood pressure after six weeks at the maximum dose.

The Data : The data analysed here will be at most two blood pressure measurements per case (one at randomisation, one after the six-week final period).

The patterns of missing data observed for this study are shown in Table 4.8.

(Note : Here, again, the data follow a nested pattern, and so iterative procedures would not be essential)

For this study, a large number of cases were removed due to side-effects :

(Placebo : $^{1}/_{6}$ of the removed cases
Captopril : $^{6}/_{8}$ of the removed cases
Hydralazine : $^{10}/_{13}$ of the removed cases
Nifedipine : $^{9}/_{9}$ of the removed cases).

Table 4.8 : Missing Data Patterns

Placebo Group

Time Point 1	Time Point 2	Number of Cases
1	1	32
1	0	6
		38

Captopril Group

Time Point 1	Time Point 2	Number of Cases
1	1	32
1	0	8
		40

Hydralazine Group

Time Point 1	Time Point 2	Number of Cases
1	1	28
1	0	13
		41

Nifedipine Group

Time Point 1	Time Point 2	Number of Cases
1	1	32
1	0	9
		41

Here it would be desirable to model the missing data mechanism, but no background information about the mechanism is available. Thus, in order to produce some kind of "answer" (more valid than if the CC or AAD approaches were used, where a MCAR setup is assumed), it was decided to proceed, with caution, using the maximum likelihood approaches.

Only the supine diastolic blood pressures (SDBP) will be studied here (although several other measurements were taken).

Comparison of the Maximum Likelihood Approach with AAD and CC

The results obtained using EM(a), AAD and CC are shown in Table 4.9 and can be summarised as follows :

(i) At the first time point, the maximum likelihood and AAD approaches agree (for reasons given in previous examples).
(ii) At the final time point, the CC and AAD approaches agree (again, for reasons given in previous examples).

Subjectively, the results are similar for the three active treatments, but these results are different from the placebo results.

Model Checking : From the transformed distance plots (Figures 4.26-4.29) there would seem to be no strong grounds on which to reject the relatively simple Multivariate Normal model in favour of the more complicated MVt or CN models. To give more objective results, MVt models with various degrees of freedom were fitted to the different groups. The maximised log-likelihoods produced are shown in Table 4.10 and Figures 4.30-4.33.

It was found that at the degrees of freedom for which the MVt log-likelihood was maximised, there was no significant difference between the MVN and MVt results for any of the groups.

Thus, a MVN model would be preferred for simplicity.

Comparison of the Groups :

(Note : The estimated means and their standard errors were practically the same for EM(a) and EM(b)).

Table 4.9 : Comparison of CC, AAD and EM(a)

Mean(s.e. of Mean)

Placebo Group

	Time Point 1	2
CC	92.9(1.9)	91.1(2.3)
AAD	93.9(1.7)	91.1(2.3)
EM(a)	93.9(1.7)	91.8(2.2)

Captopril Group

	Time Point 1	2
CC	99.9(1.8)	87.8(2.2)
AAD	98.6(1.6)	87.8(2.2)
EM(a)	98.6(1.6)	86.6(2.0)

Hydralazine Group

	Time Point 1	2
CC	98.4(1.6)	86.6(2.1)
AAD	98.6(1.3)	86.6(2.1)
EM(a)	98.6(1.3)	86.7(1.9)

Nifedipine Group

	Time Point 1	2
CC	98.0(1.8)	88.1(1.7)
AAD	97.7(1.6)	88.1(1.7)
EM(a)	97.7(1.6)	88.0(1.6)

Figure 4.26 : T.D. Plot for Placebo Group

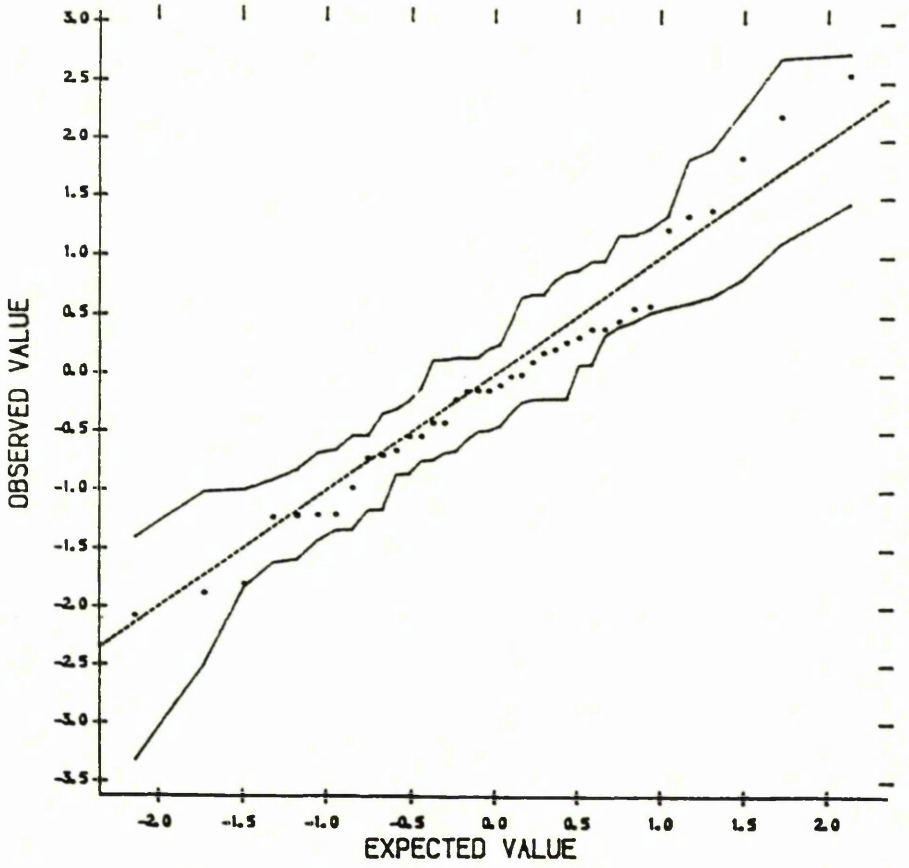


Figure 4.27 : T.D. Plot for Captopril Group

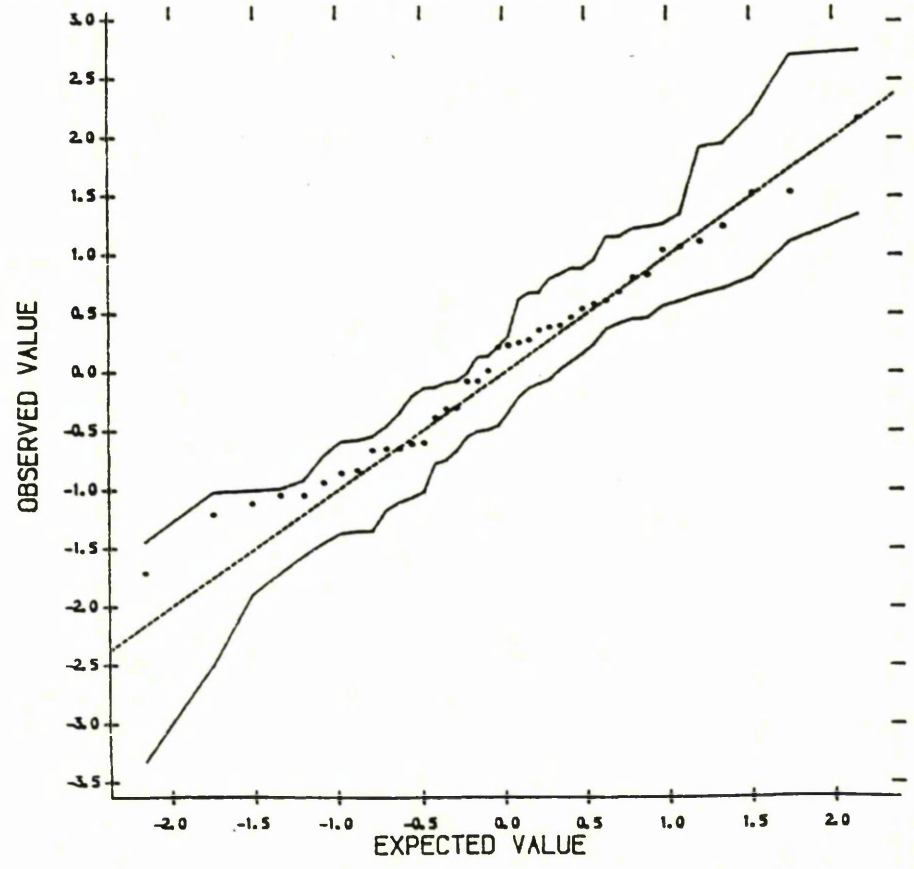


Figure 4.28 : T.D. Plot for Hydralazine Group

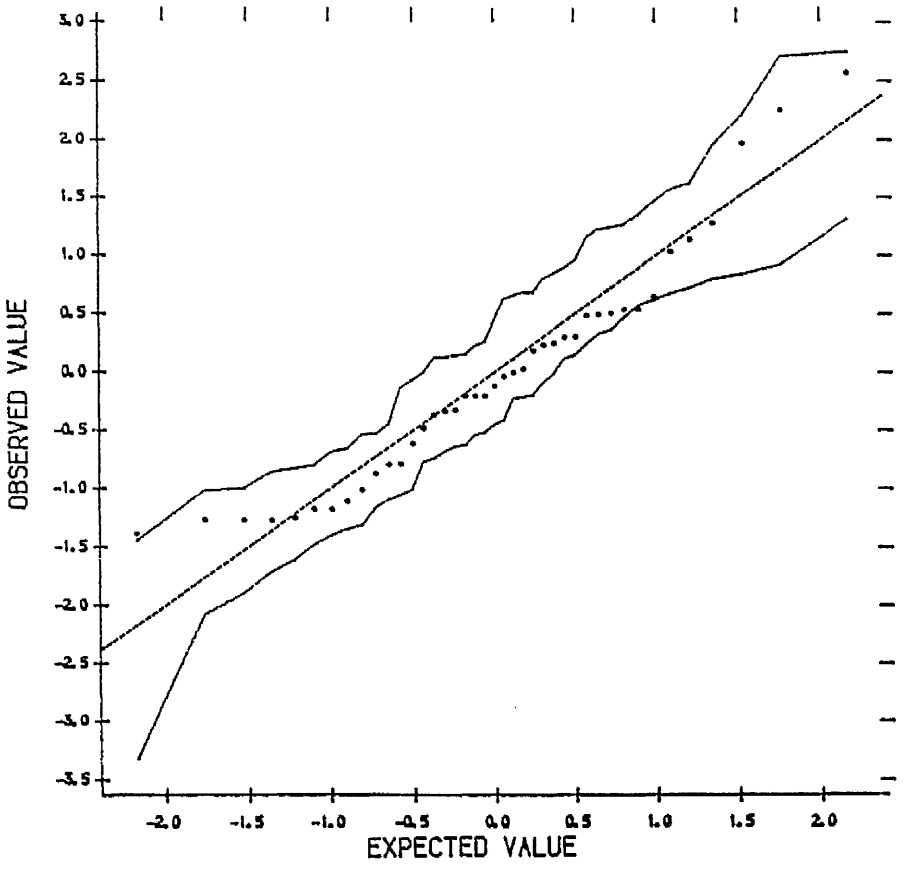


Figure 4.29 : T.D. Plot for Nifedipine Group

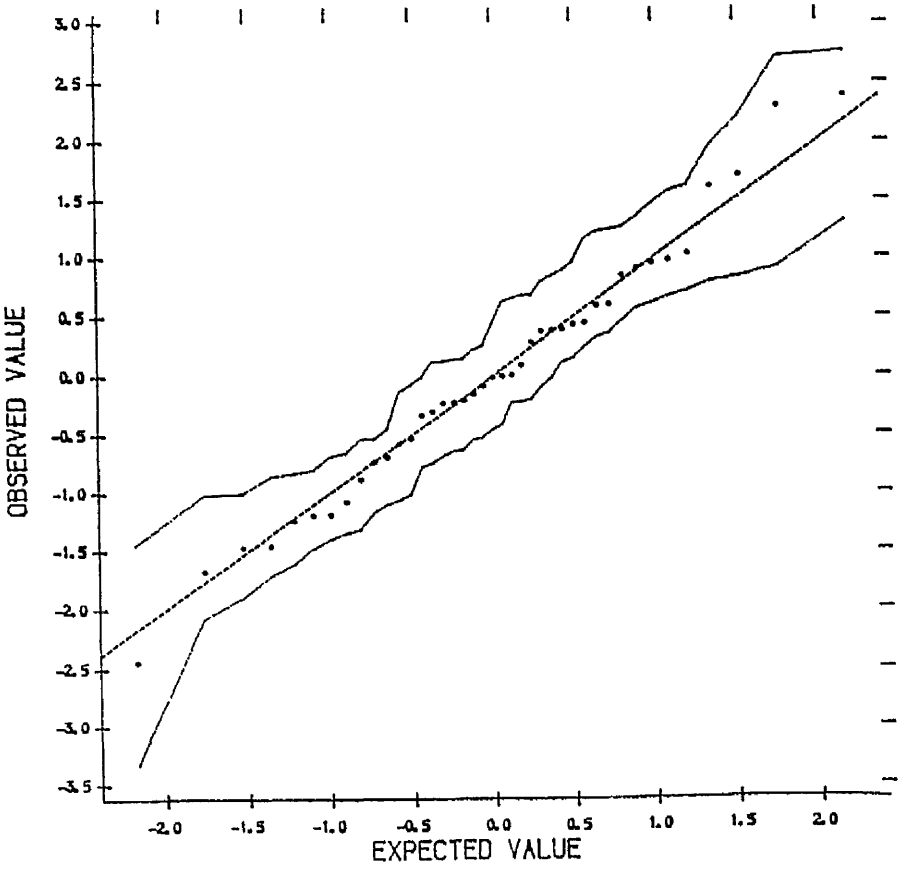


Table 4.10 : Fitting MVt Models - Maximised Log-likelihood as a Function of df

Degrees of Freedom	Placebo Group	Captopril Group	Hydralazine Group	Nifedipine Group
1	-268.1483	-276.1701	-250.7770	-272.1264
2	-263.8977	-269.7437	-245.9749	-265.9452
3	-262.7012	-267.4859	-244.5694	-263.9604
4	-262.2839	-266.3832	-244.0180	-263.0835
5	-262.1420	-265.7473	-243.7759	-262.6331
6	**-262.1122**	-265.3406	-243.6676	-262.3804
7	-262.1308	-265.0614	-243.6235	-262.2306
8	-262.1704	-264.8595	**-243.6118**	-262.1386
9	-262.2181	-264.7075	-243.6172	-262.0810
10	-262.2680	-264.5895	-243.6315	-262.0447
11	-262.3170	-264.4955	-243.6503	-262.0221
12	-262.3637	-264.4190	-243.6711	-262.0085
13	-262.4076	-264.3557	-243.6925	-262.0009
14	-262.4485	-264.3026	-243.7137	-261.9975
15	-262.4865	-264.2573	-243.7344	**-261.9968**
20	-262.6383	-264.1055	-243.8235	-262.0128
25	-262.7441	-264.0195	-243.8900	-262.0371
50	-262.9900	-263.8606	-244.0536	-262.1223
75	-263.0823	-263.8117	-244.1172	-262.1623
100	-263.1304	-263.7881	-244.1507	-262.1846
105	-263.1373	**-263.7847**	-244.1556	-262.1879

Figure 4.30 : Maximised Log-Likelihoods Produced by Fitting Various MVt Models to the Placebo Data

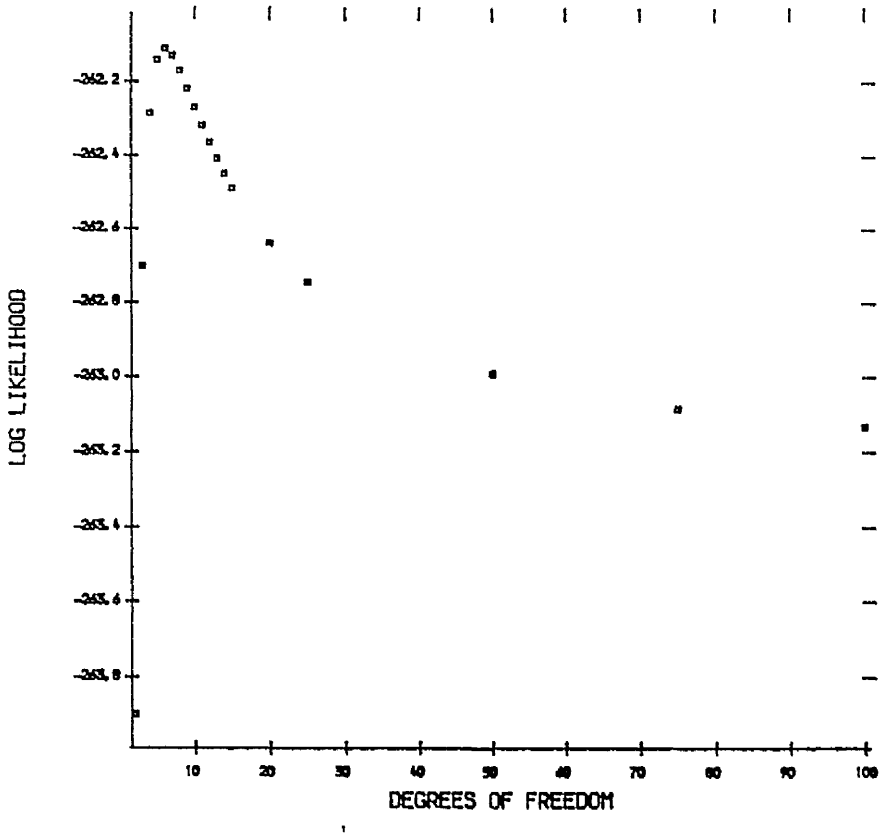


Figure 4.31 : Maximised Log-Likelihoods Produced by Fitting Various MVt Models to the Captopril Data

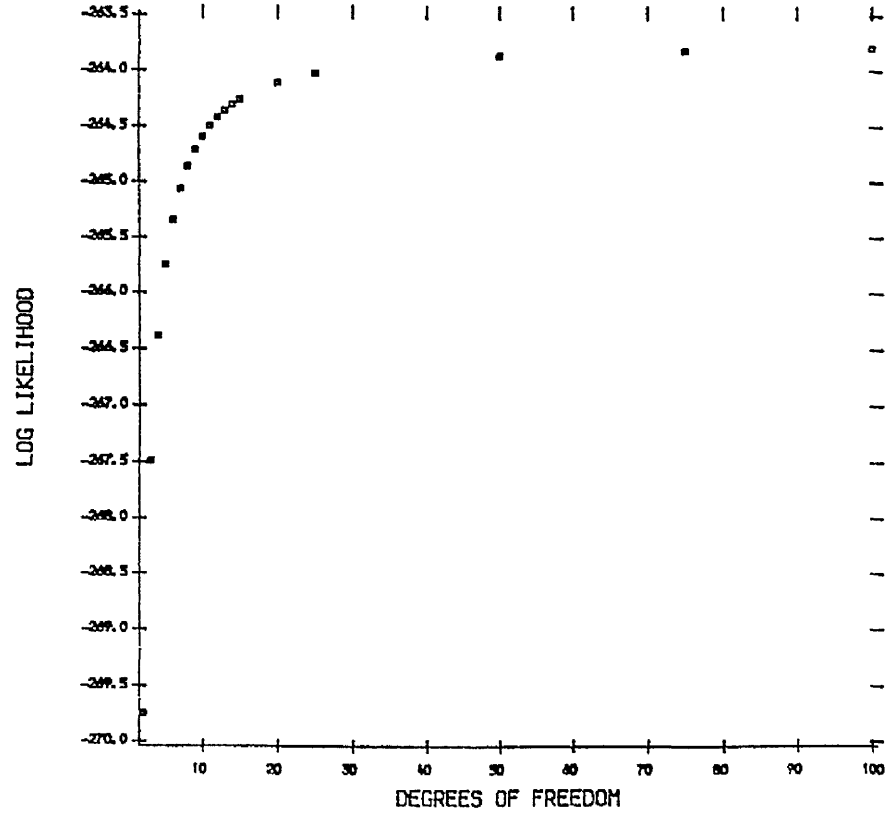


Figure 4.32 : Maximised Log-Likelihoods Produced by Fitting Various MVt Models to the Hydralazine Data

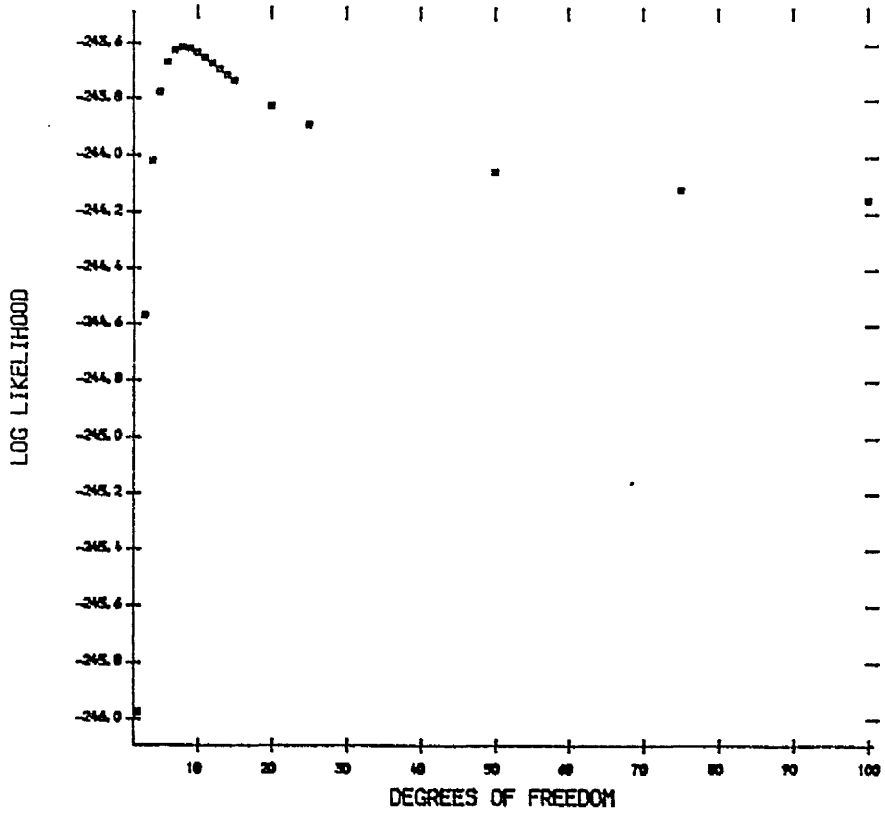
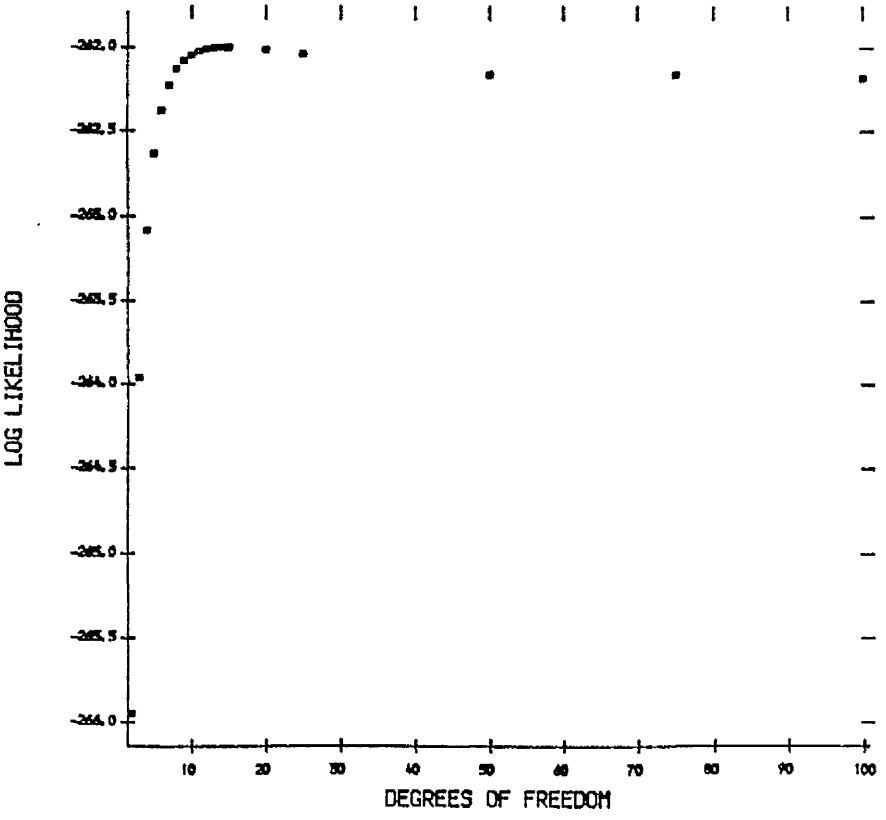


Figure 4.33 : Maximised Log-Likelihoods Produced by Fitting Various MVt Models to the Nifedipine Data



Model	Maximised log-likelihood
$\mu_1, \ldots, \mu_4 ; \Sigma_1, \ldots, \Sigma_4$	-1033.522
$\mu_1, \ldots, \mu_4 ; \Sigma$	-1037.980
$\mu ; \Sigma$	-1049.504

The results of the likelihood-ratio testing procedure were then as follows :

Stage 1 : $H_0 : \mu_i , \Sigma$ $H_1 : \mu_i , \Sigma_i$

$2 \log(\lambda) = 2(l_1 - l_0) = 8.92$

Compare this with $\chi^2({}^3/_2\, p(p+1) ; 0.95) = 16.92$

Conclusion : The covariance matrices for the four groups are not significantly different, therefore proceed to Stage 2.

Stage 2 : $H_0 : \mu , \Sigma$ $H_1 : \mu_i , \Sigma$

$2 \log(\lambda) = 2(l_1 - l_0) = 23.05$

Compare this to $\chi^2(6 ; 0.95) = 12.59$

Conclusion : There is evidence that not all four group mean vectors are equal.

Follow-up Procedure : From the means and standard errors alone, there can be seen to be differences between the placebo results and the results of the active treatments, but it is uncertain whether any of the other treatment differences are significant. To assess this, a set of simultaneous confidence intervals were produced for all pairwise differences in the blood pressure reductions in the four groups. (In these intervals, significance levels were adjusted using Bonferroni correction). The calculations are shown below. Each interval was calculated using, for $c^T\mu_i - c^T\mu_j$:

$$c^T\hat{\mu}_i - c^T\hat{\mu}_j \pm N(0,1 ; 1-{}^{0.05}/_{12}) \sqrt{\{c^T(\hat{\mathrm{cov}}(\hat{\mu}_i) + \hat{\mathrm{cov}}(\hat{\mu}_j))\, c\}}$$

$i = 1, \ldots, K-1$

$i < j$

The Results

Comparison	Point Estimate				Interval	
P vs C	-9.724	±	6.125	=	[-15.8 , -3.6]	*
P vs H	-9.851	±	6.324	=	[-16.2 , -3.5]	*
P vs N	-7.751	±	6.122	=	[-13.9 , -1.6]	*
C vs H	-0.127	±	6.319	=	[-6.4 , 6.2]	
C vs N	-1.973	±	6.117	=	[-4.1 , 8.1]	
H vs N	-2.100	±	6.316	=	[-4.2 , 8.4]	

where * denotes cases where a significant difference between two groups has been found. (In the table, the drug names have been replaced by their initial letters).

For comparison purposes, the results using a standard follow-up t-interval procedure on the complete cases are shown below (again, Bonferrini correction has been applied).

Comparison	Point Estimate				Interval	
P vs C	-10.35	±	6.39	=	[-16.7 , -4.0]	*
P vs H	-10.05	±	6.61	=	[-16.7 , -3.4]	*
P vs N	-8.06	±	6.39	=	[-14.4 , -1.7]	*
C vs H	0.30	±	6.61	=	[-6.3 , 6.9]	
C vs N	2.29	±	6.39	=	[-4.1 , 8.7]	
H vs N	1.99	±	6.61	=	[-4.6 , 8.6]	

A slight difference can be seen between the maximum-likelihood results and the complete case results in terms of the point estimates of the treatment differences.

The complete-case intervals are wider than the corresponding likelihood intervals, but the overall conclusions are unchanged, i.e. no significant differences among the active treatments, but each of the active treatments significantly different (greater blood pressure reduction) from placebo.

In this example, the method of analysis appeared to be largely irrelevant, with similar results obtained from the CC, AAD and EM(a) approaches. One would assume that this is because removal of cases due to side effects is a "missing completely at random"

process, rather than a "missing at random" process.

4.4 : Summary and Conclusions

Chapters 2-4 have dealt with a problem which commonly arises in the analysis of clinical data - that of missing data. Comparison was made of various different approaches to dealing with incomplete data, some of these ad hoc and others with more theoretical basis. The ad hoc approaches generally required that a greater number of assumptions be made about the missing data mechanism than the more theoretical approaches did (the former requiring that a "Missing Completely at Random" assumption be made, while a "Missing at Random" assumption was sufficient for the latter).

It was found that the mode of treatment of even a relatively small amount of missing data could greatly influence the results obtained (as illustrated in Example 4.1).

However, usually the analysis of clinical data is performed using one or both of the ad hoc techniques "Complete Cases" (CC) and "All Available Data" (AAD) (as described earlier). This is disturbing in that the results obtained could well be misleading. What would be proposed is that one of the theoretically-based, likelihood maximising, algorithms should be used instead, since such algorithms would be appropriate in a greater number of situations than the ad hoc approaches would be.

One drawback would be that it would be computationally prohibitive to apply the theoretically-based algorithms without the aid of a computer.

In that case, the following compromise would be proposed :

Look for two ad hoc techniques which would tend to be biased in opposite directions in the particular application of interest, thereby enclosing the "true" result. (For example, in the context of blood pressure trials, Last Value analyses would tend to overestimate the components of the mean, while All Available Data analyses would tend to underestimate them, due to the fall in blood pressure through time). Then, if there are large differences between the results obtained using the two ad hoc techniques, at <u>that</u> stage implement an approach such as EM(a).

Obviously this is not ideal, since information is being lost by using AAD or LV, but hopefully is a reasonable solution to a

difficult problem.

It should be noted that situations can arise where neither a MCAR nor a MAR assumption is appropriate, for example when the data take the form of survival times. However, unfortunately it is not possible to test the MCAR and MAR assumptions, since to test these assumptions, one would have to analyse the values which would have been observed - this is not possible.

Instead, it is necessary to go back to the context from which the data arose.

In the example comparing Ketanserin with Metoprolol, it could be seen from the design of the study that a large proportion of the missing data could be attributed to a MAR mechanism, while for the Third Drug Study, data were more likely to be missing due to a MCAR mechanism.

Section 3.3 progressed to the K-sample problem and attempted to provide an answer to the problem of how best to handle incomplete data from multiple groups.

The likelihood approaches from the one-sample problem were extended to allow estimation of the parameters from more than one group, while assuming a common covariance matrix for these groups.

This allowed emulation of the techniques for comparing groups usually applied only to complete data.

In addition, large-sample covariance matrices for the estimated mean vectors were provided, allowing approximate confidence intervals for contrasts of the means (or linear combinations of the means) to be produced.

PART II : Order Restricted Inference

Chapter 5 : Background to the Problem

5.1 : Introduction and Literature Review

As mentioned in the Chapter 1, in the analysis of clinical data, there will often exist some prior information as regards the magnitude of the expected responses under different treatments, enabling an ordering to be imposed on the underlying group mean responses during analysis.

For example, in the early stages of development of a drug, it is common to assess its effect by comparing the magnitude of responses obtained under different doses of the drug to those obtained under placebo.

Clearly, if the drug has no effect, then the responses in the different groups would be expected to be similar.

However, if the drug does have some effect, then there is a natural ordering in the magnitude of responses which would be expected, in that the placebo would be expected to produce the least response and the highest dose of drug would be expected to produce the greatest response, with intermediate doses expected to produce intermediate, ordered, levels of response.

Imagine, for example, that in the study of an antihypertensive agent, X, individuals are randomised to one of three groups :

Group 1 : Placebo
Group 2 : Low dose of X
Group 3 : High dose of X .

Let the response of interest be the fall in the Mean Arterial Pressure during eight weeks of treatment.

Letting μ_i represent the underlying mean response for Group i, (i= 1, 2, 3), it would be expected that there would be an ordering : $\mu_1 \leqq \mu_2 \leqq \mu_3$.

In this context, when testing for the presence of a treatment effect, instead of testing the Null Hypothesis,

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

against the usual alternative hypothesis of

H_1 : Not all of the μ_i are equal ,

it would utilise more prior knowledge, and give a more powerful test, if the same null hypothesis was tested against the alternative hypothesis

H_A : $\mu_1 \leq \mu_2 \leq \mu_3$ where at least one of the inequalities is strict.

Since the objective here is to increase the statistical sensitivity of analyses performed, by incorporation of what useful background information we have, it is natural to consider in the same light other means of improving the sensitivity, such as by incorporating useful covariates into analyses and/or by making appropriate distributional assumptions about the responses under study.

The Notation

(Y_{ij}, X_{ij}) = (Response , Covariate) pair for Subject j of Group i, $(i=1,\ldots,k$ $j=1,\ldots,n_i)$

n_i = Number of cases in Group i $(i=1,\ldots,k)$

μ_i = Mean response for Group i $(i=1,\ldots,k)$

5.2 : Statistical Approaches

5.2.1 : Review of the Existing Literature

Over the last forty years, much work has been carried out in the area of "Order Restricted Inference", (i.e testing hypotheses of form H_0 vs H_A above).

Two methods which have essentially been adopted as standard, and so have been used consequently as a comparison to other methods emerging, are those proposed by Bartholomew(1959) and Jonkheere(1954). (The method of Jonkheere was also proposed independently by Terpstra in 1952.)

For the situation where $F_i(Y)$ represents the (continuous) cumulative distribution function of Group i, and where interest

lies in testing whether

$$F_1(Y) \leq \ldots \leq F_k(Y) \quad ,$$

Jonkheere adopted a non-parametric approach to the problem, proposing the test statistic :

$$J = 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} p_{ij} - \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_i n_j$$

$$\text{where} \quad p_{ij} = \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} p_{irjs}$$

$$\text{with} \quad p_{irjs} = \begin{cases} 1 & \text{if } Y_{ir} < Y_{js} \\ 0 & \text{if } Y_{ir} \geq Y_{js} \end{cases} \quad .$$

Under the null hypothesis, p_{irjs} would have equal probability of taking value 0 and taking value 1, so that J would have expectation zero.

With the group means ordered as in the alternative hypothesis, p_{irjs} would have a higher probability of taking value 1, so that J would have expectation greater than zero. Jonkheere established that under H_0, J could be represented, asymptotically, by a Normal random variable, with mean zero and variance

$$1/18 \left(N^2 (2N + 3) - \sum_{i=1}^{k} n_i^2 (2n_i + 3) \right)$$

$$\text{where} \quad N = \sum_{i=1}^{k} n_i \quad .$$

If we let MW_{ij} represent the Mann-Whitney(1947) statistic for comparing samples i and j, then J can be written as :

$$J = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} MW_{ij}$$

$$\text{where } MW_{ij} = \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} I(Y_{ir}, Y_{js})$$

with I(u,v) an indicator of the event $u < v$.

Bartholomew's method was constructed under an assumption of Normality in the responses for the populations under study.

Let Y_{ij} represent the response for Subject j from Group i, and let $Y_{ij} \sim N(\mu_i, \sigma^2)$.

To form Bartholomew's test statistic, it is necessary to calculate the minimised Sum of Squares for the responses about their group means as estimated under the alternative hypothesis, thus it is necessary to calculate maximum likelihood estimates for the k population means, $\mu_1, \ldots, \mu_k$, under the ordering restrictions imposed by the alternative hypothesis. These estimates are found by carrying out a technique known as 'Isotonic Regression' on the k sample means weighting by the group sample sizes.

Several algorithms for Isotonic Regression were described by Barlow et al in 1972. The algorithm used for the present work was the 'Up and Down Blocks Algorithm' (proposed by Kruskal(1964)), as described in Appendix 1.

Bartholomew's test statistic was then defined as :

$$\bar{E}_k^2 = \sum_{i=1}^{k} n_i(\hat{\mu}_i^* - \hat{\mu})^2 \Big/ \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\mu})^2$$

where $\hat{\mu}_i^*$ is the isotonic estimate of the i^{th} population mean, and

$$\hat{\mu} = \sum_{i=1}^{k} n_i\bar{Y}_i \Big/ \sum_{i=1}^{k} n_i$$

is the maximum likelihood estimate estimate for the common population mean under H_0. The null distribution of the test statistic is given by:

$$P(\bar{E}_k^2 \geq c \mid H_0) = \sum_{\ell=1}^{k} P(\ell,k;w)\, P(\beta_{(\ell-1)/2,(N-\ell)/2} \geq c)$$

where $P(\ell,k;w)$ is the probability that, under H_0, and for weights w_i, the isotonic regression on the k sample means yields exactly ℓ distinct values, and $\beta_{a,b}$ denotes the Beta distribution with a and b degrees of freedom.

Note : To calculate the value of c corresponding to a given significance level, it is necessary to use numerical methods (e.g. bisection techniques). However, some limited tables of critical values are available for the case of equal sample sizes in, for example, Bartholomew et al (1972)).

In addition to Jonkheere's test, other distribution-free approaches have been suggested. Chacko(1963) proposed a technique (extended by Shorack(1967)) whereby observations were replaced by their rank (R_{ij} corresponding to Y_{ij}) in the order statistic for all of the observations. Isotonic regression was then carried out on the mean group ranks, $\bar{R}_i$ to produce $\bar{R}_i^*$, their corresponding isotonic estimates. The test statistic was then defined as

$$H_k^{-2} = {}^{12}/_{n(n+1)} \sum_{i=1}^{k} n_i (\bar{R}_i^* - {}^{(N+1)}/_2)^2 \quad ,$$

with null distribution given by

$$P(H_k^{-2} \geq c \mid H_0) = \sum_{l=1}^{k} P(l,k) \; P(\chi^2_{l-1} \geq c)$$

where $P(l,k)$ is the equal-weights equivalent of $P(l,k;w)$.

Doksum(1967) and Hollander(1967) both proposed tests based on scores, the scores measuring the "magnitude" of difference between given pairs of groups. It was noted that many common rank-based statistics could be written in the form

$\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} S_{ij}$ where S_{ij} is some measure of the difference between groups i and j.

For example, if some score is defined as taking value 1 if $Y_{ir} \leqq Y_{js}$ and value 0 otherwise, and then these quantities are summed over r and s, to produce S_{ij}, the statistic defined in the previous paragraph would be equivalent to Jonkheere's J statistic.

Hollander and Doksum modified the form of S_{ij} to take account of the magnitude of the difference between the pairs of groups. Their test statistics were based upon ranking the differences between the pairs of observations in each pair of groups, and constructing

$$S_{ij} = \sum_{h=1}^{n} I(Y_{hi} \angle Y_{hj}) \; R_{ij}(h)$$

where $R_{ij}(h)$ is the rank of $|Y_{hi} - Y_{hj}|$ among all such quantities for fixed i and j,

and $I(Y_{hi} \angle Y_{hj})$ is an indicator of the event $Y_{hi} \angle Y_{hj}$

Hollander's statistic was then defined as :

$$Y_1 = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} S_{ij}$$

while Doksum's statistic was given by the related statistic

$$Y_2 = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} S_{i.} - S_{j.}$$

However, Y_1 and Y_2 are not distribution-free, even asymptotically, since the correlation between any pair of observations would depend on the form of the underlying distribution, F(y). It is possible, however, to use techniques proposed by Lehmann(1964), whereby the correlation coefficients are estimated from the data, to produce asymptotically distribution-free tests.

Several papers have been published carrying comparisons of the aforementioned test statistics.

Bartholomew(1961) compared the asymptotic power functions of $\bar{E}_k^2$ and J with a classical F-test (i.e. prior information with respect to group ordering ignored), for k = 3 and k = 4, and for :

Case 1 : Equal-sized spacings of the true population means

Case 2 : All-except-one of the population means equal

In each case the responses were assumed to come from independent Normal populations.

His conclusions were that for both k = 3 and k = 4, and for both Case 1 and Case 2, there was little difference between $\bar{E}_k^2$ and J, with J slightly more powerful in Case 1 and $\bar{E}_k^2$ slightly more powerful in Case 2. (However, both of these tests were substantially more powerful than a classical F-test.

For $k \geq 4$, J would still be more powerful than $\bar{E}_k^2$ for Case 1, but much less powerful for Case 2.

Puri(1965), after proposing a family of test statistics, V, encompassing both the Normal Scores test ($V(\Phi)$) and Jonkheere's test, J, went on to make asymptotic power comparisons of $V(\Phi)$, J, $\bar{E}_k^2$ and F (the classical F-test), under an assumption of Normality in the responses. Again, cases 1 and 2 as defined above were considered. His results were as follows :

(i) All three tests incorporating the ordering information always surpassed the classical F-test.

(ii) The power of all three ordered tests were lower for Case 2 than for Case 1, where compared here were tests with the same value for the quantity, Δ, defined by :

$$\Delta^2 = \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2$$

(iii) The Normal Scores test was more powerful than both J and $\bar{E}_k^2$ in Case 1, and more powerful than J but less powerful than $\bar{E}_k^2$ for Case 2.

Magel(1983) performed a Monte Carlo simulation study to compare the Kruskal-Wallis test (1-way non-parametric Analysis of Variance), with J and Chacko's $\bar{H}_k^2$ statistic. The response distributions considered were Uniform, Normal and Cauchy. In each situation, the conclusions were that for Case 1 (defined above), J was most powerful, and for Case 2, $\bar{H}_k^2$ was most powerful, with the Kruskal-Wallis test performing the worst in each case.

5.2.2 : Covariates

Returning to a point made in Chapter 1, it is common to have available covariate information corresponding to the response for each individual, and it would often be desirable to adjust, in some sense, for this covariate information when performing any analyses of the responses. The reasons for this are twofold. Firstly, the adjustment for appropriate covariates can increase the sensitivity of statistical analyses performed, by reducing the response variability. Secondly, by incorporating covariates into analyses, one can nullify the effects of imbalances in the covariate distributions for different groups (although such differences should be small for randomised studies).

In the absence of any prior information with respect to ordering, there exist standard techniques for comparing treatments while allowing for covariate information, the appropriate form of analysis being chosen dependent on whether certain assumptions could be made about the data.

For a standard Analysis of Covariance based on the model :

$$Y_{ij} = \alpha_i + \beta (X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}$$

$$\text{where } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \quad ,$$

the necessary assumptions would be :

(1) Normality of the conditional responses, given the covariate values, i.e. $Y_{ij} | X_{ij}$.
(2) Equality of the conditional Y-variances for the different groups.
(3) Linearity of the within-group regressions
(4) Equal regression slopes for each of the groups concerned.
(5) Random assignment of individuals to the treatment groups.
(6) Error-free measurement of the covariate values
(7) Fixed treatment levels, i.e. not randomly selected from the population of all treatment levels.
(8) Independence of the $\{ \varepsilon_{ij} \}$.

In addition, interpretation of results is eased if it can be assumed that the treatment levels and the covariates are unrelated.

In a standard parametric Analysis of Covariance, the test statistic considered is

$$F_2 = \frac{SSAT / (K-1)}{SSres_w / \left(\sum_{i=1}^{k} n_i - K - C \right)}$$

where K = Number of groups

and C = Number of covariates

$SSres_w$ = Within-groups Residual Sum of Squares

$$= \sum_{i=1}^{k} Syy_i - \sum_{i=1}^{k} Sxy_i / \sum_{i=1}^{k} Sxx_i$$

$SSres_t$ = Total Residual Sum of Squares

$$= Syy - Sxy / Sxx$$

$$SSAT = SSres_t - SSres_w$$

In these expressions,

$$Sxy_i = \sum_{j=1}^{n_i} X_{ij} Y_{ij} - n_i \bar{X}_{i.} \bar{Y}_{i.}$$

$$Sxy = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}Y_{ij} - N\bar{X}_{..}\bar{Y}_{..}$$

In each case, a dot and a bar denote averaging over all possible values of that particular subscript. The remainder of terms in the Test Statistic defined follow logically.

Under a Null Hypothesis of

H_0 : " No difference between the treatment effects ",

i.e equality of the group means when adjusted for the covariates,

$$F_2 \sim F\left(K-1, \sum_{i=1}^{k} n_i - k - C\right)$$

If a situation arose where it was not possible to make any/all of the assumptions (1) - (3), it would be advisable to adopt a distribution-free approach to the problem.

One method based on ranks was proposed by Quade(1967). For the moment, only the one-covariate case will be detailed, although generalisations for the case of more than one covariate are straightforward.

For this one-covariate case, the procedure was clearly laid-out by Huitema(1980) as follows :

(i) Rank separately the responses, Y_{ij}, and the covariates, X_{ij}.

(ii) Convert these ranks into "deviation ranks", by subtraction of the mean rank, i.e take

$$x_{rank}(i,j) = X_{rank}(i,j) - \bar{X}_{rank}$$
$$y_{rank}(i,j) = Y_{rank}(i,j) - \bar{Y}_{rank}$$

(iii) Regress y_{rank} on x_{rank} . The slope parameter yields the value of Spearman's rank correlation coefficient, r_s.

(iv) Use the x_{rank} to predict the corresponding deviation rank for y, $\hat{y}_{rank}$:

$$\hat{y}_{rank}(i,j) = r_s \cdot x_{rank}(i,j)$$

(v) Calculate a "residual", z_{ij}, for each individual using

$$z_{ij} = y_{rank}(i,j) - \hat{y}_{rank}(i,j)$$

(vi) A valid test for equality of the conditional distributions of Y given X can then be performed by carrying out a standard parametric analysis of Variance on the "residuals", z_{ij}

Little work has been published about methods appropriate where there is both prior information about the ordering of the population means and also covariate information available.

Quade(1982) proposed a method for non-parametric Analysis of Covariance which was extended by Marcus and Genizi(1987) for the situation where ordering information is available. In Quade's paper, the reasoning was as follows : If the null hypothesis was true, then the samples from different groups could be pooled. Therefore, to test whether the populations were identical, one could pool the samples to determine the relationship of Y and X, then compare each true response Y_{ij} with the value which would be predicted for Y_{ij} from the relevant covariate, X_{ij} and the fitted regression line for the pooled data.

A score could then be assigned, its value dependent on whether the observed response was larger or smaller than its predicted value. A one-way Analysis of Variance could then be performed on these scores. The method proposed by Quade for estimating the responses from their covariates involved a technique called "Caliper Matching", where two individuals are described as 'matched' if their covariates are, at most, a distance ε apart ($D(X_{ij},X_{i'j'}) \leq \varepsilon$). This could be interpreted as identifying cases whose background information was "similar" in some sense.

Quade proposed estimating the response for each case by the average of the responses of all cases matched-by-covariate with that case.

$$\hat{Y}_{ij} = \sum_{i'=1}^{k} \sum_{j'=1}^{n_{i'}} Y_{ij}\, I(D(X_{ij},X_{i'j'}) \leq \varepsilon) \;/\; M_{ij}$$

where $I(D(u,v) \leq \varepsilon)$ is an indicator of the event $D(u,v) \leq \varepsilon$ and Mij is the number of cases matched-by-covariate with case (i,j). The "score" to be used in the one-way analysis of variance would then be the difference between $\hat{Y}_{ij}$ and Y_{ij}.

Marcus and Genizi's application and adaptation of Quade's techniques were as follows :

Let $\theta_1, \theta_2, \ldots, \theta_k$ denote the k treatment effects, where interest lies in testing the hypotheses

$H_0 : \theta_1 = \theta_2 = \theta_3$ against the alternative

$H_A : \theta_1 \leq \theta_2 \leq \theta_3$ where at least one inequality is strict.

Define $M(X_{ij},X_{i'j'})$ as the matching funtion for cases (i,j)

and (i',j'). In Marcus and Genizi's paper, $M(X_{ij},X_{i'j'})$ took the simple form of a binary variable taking a value zero if these cases were not matched ($D(X_{ij},X_{i'j'}) \succ \varepsilon$) , and taking a value 1 if they were matched ($D(X_{ij},X_{i'j'}) \leqq \varepsilon$).

In Quade's 1982 paper, another, more complicated form of matching function was proposed, although this possibility was not investigated by Marcus and Genizi. This 'other' form of matching function was defined as

$$M(X_{ij}, X_{i'j'}) = \exp (- d)$$
$$\text{where } d = D (X_{ij}, X_{i'j'})$$

Clearly, a more flexible form of this function to use would be

$$M(X_{ij}, X_{i'j'}) = \exp (- k\, d)$$

Here, $M(X_{ij},X_{i'j'})$ would take values of 0 and 1 at the extremes of $d = \infty$ and $d = 0$, respectively, with a continuous exponential decay function between these extremes (rather than a discontinuous step function).

$$\text{Let} \quad M = \sum_{i=1}^{k-1} \sum_{i'=i+1}^{k} \sum_{j,j'} M (X_{ij},X_{i'j'})$$

$$\text{and} \quad L = \sum_{i=1}^{k-1} \sum_{i'=i+1}^{k} \sum_{j,j'} M (X_{ij},X_{i'j'}) \cdot I (Y_{ij}, Y_{i'j'})$$

where I(u,v) is an indicator of the event $u \prec v$.

For the case of a binary matching function, M could be interpreted as the "Total number of matched pairs", and L could be interpreted as the "Number of these pairs which are 'correctly' ordered in terms of the alternative hypothesis".

Marcus and Genizi proposed a test statistic

$$Q = (W - 0.5) / s_w$$

where $W = L / M$, and s_w^2 is a conditional estimate of the variance of W (conditional on the continuous covariate, X).

It was suggested that, due to the asymptotic Normality of Q, an asymptotic α-level test of H_0 against H_A could be performed by referring Q to the $(1-\alpha)$ quantile of the Standard Normal distribution. To find s_w^2, Marcus and Genizi extended the Method of Components of Sen(1960) as in Quade(1974), showing that

$$M^2 s_w^2 = \sum_{i=1}^{k} \left[\sum_j L_{ij}^2 - \frac{L_{i0}^2}{n_i} - 2W \left[\sum_j L_{ij}M_{ij} - \frac{L_{i0}M_{i0}}{n_i} \right] + W^2 \left[\sum_j M_{ij}^2 - \frac{M_{i0}^2}{n_i} \right] \right]$$

where

$$L_{ij} = \sum_{i=1}^{k-1} \sum_{i'=i+1}^{k} \sum_{j,j'} M(X_{ij},X_{i'j'})I(Y_{ij},Y_{i'j'}) + \sum_{i'=1}^{k-1} \sum_{i=i'+1}^{k} \sum_{j,j'} M(X_{ij},X_{i'j'})\, I(Y_{i'j'},Y_{ij})$$

$$M_{ij} = \sum_{i'=1}^{i-1} \sum_{j'=1}^{n_{i'}} M(X_{ij},X_{i'j'}) + \sum_{i'=i+1}^{k} \sum_{j'=1}^{n_{i'}} M(X_{ij},X_{i'j'})$$

$$L_{i0} = \sum_j L_{ij} \qquad M_{i0} = \sum_j M_{ij}$$

5.2.3 : Discussion

From the papers discussed, several points emerged :
When looking at response data there will often exist potential sources of useful information, where 'appropriate' use of such information would lead to more sensitive analyses of the responses themselves.

e.g (i) Covariate Information corresponding to each response
(ii) Normality Assumptions about these responses
(iii) Prior Ordering Information about the group means

However, no papers were available to compare statistical tests incorporating varying amounts of the prior information as defined in (i) - (iii) above.

In addition, it became evident that there were no tests available to incorporate information from all three of the sources simultaneously, i.e there was not available a test incorporating a Normality assumption about the (conditional) responses, prior ordering information about the group means and

also covariate information.

If the ordering of responses was due to differing doses of the same treatment in the groups, one possibility might be to include the dose of drug as a covariate in the analysis of the responses. However, a difficulty arises in that the dose-response relationship is generally a non-linear one. Typical dose-response curves tend to be of sigmoidal form. This means that as dose increases, there is a lower rate of change in the responses at low and high doses than there is within some central dose-range. Thus, with the linearity assumption violated, it would not be appropriate simply to incorporate dose as a covariate.

Because of the 'gaps' in the existing literature, the following objectives emerged.

5.2.4 : The Objectives

(1) To produce a method which incorporates information from all three sources identified in the previous section.

(2) To compare this derived test to existing tests which incorporate varying amounts of prior knowledge about ordering, covariates, and distributional form for the responses, looking at the tests' performances when :
   (a) The magnitude of the group spacings are varied
   (b) The relative spacings of the groups are varied
   (c) The correlation between the responses and covariates is varied
   (d) The error distribution is non-Normal

Note : in the simulations to follow, the number of observations per group will be fixed at twenty. It is natural to assume that the pattern of results for other choices of sample size would be similar (although obviously not identical). One change which would be expected as the sample size increased would be that the performance of the non-parametric tests would be expected to improve, so as to become more similar to that of their corresponding Normal test.

Clearly there are eight possible combinations of the three two-level factors of interest :

(i) Covariates : Incorporated / Ignored

(ii) Normality of the (conditional) responses : Assumed / Not assumed

(iii) Ordering Information : Incorporated / Ignored

- each one of these eight combinations defining a certain type of test (see Table 5.1). Certain of these tests are well-known (e.g Analysis of Variance for the combination "Covariates ignored, Normality assumed, Ordering Information ignored") while others have either been described in the literature review or will now be described.

5.2.5 : The Tests

Test 1 : No covariates ; Normality assumed ; No Ordering

- Analysis of Variance on the response data.

Test 2 : Covariates used ; Normality assumed ; No Ordering

- Analysis of Covariance on the response data.

Test 3 : No Covariates ; Normality not assumed ; No Ordering

- Kruskal-Wallis Rank Analysis of Variance on the responses

Test 4 : Covariates used ; Normality not assumed ; No Ordering

- Rank Analysis of Covariance on the responses (Huitema(1980))

Test 5 : No Covariates ; Normality assumed ; Ordering Used

- Bartholomew's $\bar{E}_k^2$ test on Y_{ij} (Bartholomew(1959))

Test 6 : Covariates used ; Normality assumed ; Ordering used

- A proposed modification of Bartholomew's test (Test 5) to allow incorporation of covariate information.
- Strictly speaking, it was the responses, rather than the form of $\bar{E}_k^2$ which were modified. Let (X_{ij}, Y_{ij}) represent the (Covariate , Response) pair for case j in group i $(i=1,\ldots,k\ ,\ j=1,\ldots,n_i)$, and let $\bar{X}_{..}$ denote the grand mean for all of the covariates , i.e.

Table 5.1 : The Tests

FACTOR (i)	FACTOR (ii)	FACTOR (iii)		TEST PERFORMED
x	✓	x	1	Analysis of Variance on the Y_{ij}
✓	✓	x	2	Analysis of Covariance on Y_{ij} vs X_{ij}
x	x	x	3	Kruskal - Wallis Rank Analysis of Variance on Y_{ij}
✓	x	x	4	Rank Analysis of Covariance on Y_{ij} vs X_{ij} (cf Huitema(1980))
x	✓	✓	5	Bartholomew's $\bar{E}_k^2$ test on Y_{ij} (cf Bartholomew(1959))
✓	✓	✓	6	Proposed test ($\bar{E}_{adj}^2$) (See following section for details)
x	x	✓	7	Jonkheere's test, J (cf Jonkheere(1954))
✓	x	✓	8	Marcus and Genizi's test, based on Q (cf Marcus and Genizi(1987))

Where ✓ denotes Assumed / Incorporated
and x denotes Not assumed / Ignored

Recall that the factors were as shown :

(i) Covariates

(ii) Normality of the (conditional) responses

(iii) Ordering Information

$$\bar{X}_{..} = N^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

Each response, Y_{ij}, was adjusted using its own covariate value, and the regression of Y against X . Transformation of the responses was made, using

$$Y_{ij}' = Y_{ij} - \hat{\beta} (X_{ij} - \bar{X}_{..})$$

where $\hat{\beta}$ is the maximum likelihood estimate for the gradient of the linear regression of Y against X, assuming the same slope parameter for each group. In addition to reducing the variability in the responses, this adjustment procedure would also correct for differences among the covariate means for the groups, although such differences should be small for randomised studies.
Bartholomew's $\bar{E}_k^2$ test was then simply performed on the adjusted responses, Y_{ij}', rather than on the original responses, Y_{ij}.

Test 7 : No Covariates ; Normality not assumed ; Ordering used

- Jonkheere's test, J, on the responses (Jonkheere(1954))

Test 8 : Covariates used ; Normality not assumed ; Ordering used

- Marcus & Genizi's test, Q (Marcus & Genizi(1987))

Chapter 6 : A Comparative Study Based on Simulation

6.1 : Scope of the Simulations

In this chapter, data will be simulated from linear regression models of the form

$$Y_{ij} = \alpha_i + \beta X_{ij} + \varepsilon_{ij}$$
$$\text{where } \varepsilon_{ij} \sim N(0, \sigma_y^2)$$
$$\text{and } X_{ij} \sim N(\mu_x, \sigma_x^2)$$
$$i = 1, \ldots, k$$
$$j = 1, \ldots, n_i$$

In the simulations, certain parameters will be fixed, namely

$$\mu_x = 115 \quad , \quad \sigma_x^2 = 25 \quad , \quad \sigma_y^2 = 45$$

Although the chosen family of models appears to be very specific and limited, a few important points should be made :

(i) The choice of μ_x is irrelevant - a different choice of μ_x would alter only the location of the responses, not the conclusions.

(ii) The only critical model choices to be made are
   (a) The correlation, ρ, between the response and covariate variables
   (b) The number of observations in the groups, n_i $(i = 1, \ldots, k)$
   (c) The differences between the intercepts in relation to σ_y (the marginal standard deviation for Y)
   (d) The covariate variance, σ_x^2

In some of the later stages of the chapter, where stated, the error distributions will be of certain given non-Normal forms.

6.2 : Refinement of the Models

6.2.1 : Evaluation of the Marcus & Genizi Test

Before progressing to compare tests 1-8 above, it was necessary first to decide on a suitable form for the matching function in Test 8. Recall that Marcus and Genizi's test statistic, Q was defined as

$$Q = (W - 0.5) / s_W$$

$$\text{where } W = \frac{\sum_{i=1}^{k-1} \sum_{i'=i+1}^{k} \sum_{j,j'} M(X_{ij},X_{i'j'})\, I(Y_{ij},Y_{i'j'})}{\sum_{i=1}^{k-1} \sum_{i'=i+1}^{k} \sum_{j,j'} M(X_{ij},X_{i'j'})}$$

$M(X_{ij},X_{i'j'})$ was the function defining the matching relationship of cases (i,j) and (i',j'), the form of which had not yet been chosen.

The Choice of Matching Function

Two general forms of matching function were considered

(A) Caliper Matching, where the matching function was defined as

$$M(X_{ij},X_{i'j'}) = \begin{cases} 0 & \text{if } D(X_{ij},X_{i'j'}) > c \,.\, IQR \\ 1 & \text{if } D(X_{ij},X_{i'j'}) \leq c \,.\, IQR \end{cases}$$

(This was studied for $c = 0.05, 0.25, 0.5, 0.625, 0.75, 1.5, 2.5, \infty$)

(B) Continuous Matching Function, where the matching function was defined as

$$M(X_{ij},X_{i'j'}) = \exp\left[- \frac{k \,.\, D(X_{ij},X_{i'j'})}{IQR} \right]$$

(This was studied for $k = 20, 7.5, 5.0, 2.5, 0.5$)

In each case, $D(X_{ij},X_{i'j'})$ is the 'distance' between the covariates, i.e $|X_{ij} - X_{i'j'}|$, and IQR denotes the interquartile range for the covariates.

In both (A) and (B), the justification for the introduction of the interquartile range was to eliminate the dependence on the units of measurement. Otherwise, a different choice of units for the same data would lead to a different choice of c or k.

Clearly, the optimal choice of c or k will depend on the number of observations in the groups. To exploit the ordering properties of matched observations, it is desirable to have a large number of these matched observations. So if the sample

sizes were small, leading to very sparse data, the optimal caliper width would be wider, in order to take in more pairs (i.e c would be larger), and the optimal choice of decay function would be one with a more gradual fall (i.e k would be smaller), than they would be if the data was more abundant. Similarly, increasing the covariate variability would lead to larger optimal values of c and smaller optimal values of k.

The choice of matching function is analogous to the choice of smoothing parameter in the problem of density estimation : in both cases, it is desirable to perform greater smoothing if fewer observations are available (See Silverman(1986)).

6.2.2 : Optimisation of the Marcus and Genizi Test

A simulation study was carried out to assess and compare the performance of Marcus and Genizi's test for the several choices of c and k defined above. The study consisted of two parts :

(A) For a given model, simulate data sets a large number of times (20000) and calculate statistics Q and W.

(a) Assess the Normality of these statistics by way of their samples skewness and kurtosis and by histograms of the test statistic values.

(b) Assess the Standard Normal approximation for Q by studying its sample variance, comparing the histograms of Q to those expected from a Standard Normal distribution, and by finding the percentage of the values of Q which would be rejected if they were compared to the 95th and 90th percentiles of a Standard Normal distribution.

(c) Estimate the critical values for Q and W directly from their simulated distributions.

(B) Use the estimated critical values from (A) to carry out simulations to assess and compare the performance of the different test-options.

Model Specifications for (A)

$$\mu_x = 115$$
$$\sigma_x^2 = 25$$
$$\sigma_y^2 = 45$$
$$\rho^2 = 0.80$$
$$\alpha_1 = \alpha_2 = \alpha_3 = -15$$
$$\beta = \rho\, \sigma_y / \sigma_x$$
$$n_i = 20 \qquad i=1, \ldots, k$$
$$k = 3$$

Here the choices of σ_x^2, ρ, and μ_x were made to mimic the results seen in a typical clinical trial, where Y was the final systolic blood pressure, and X was the inital (baseline) systolic blood pressure.

The Results for (A)

The sample skewness and kurtosis for the simulated distributions are shown in Tables 6.1 & 6.2 together with sample means and variances. It can be seen that for W, the skewness and kurtosis are not inconsistent with what would be expected if the distributions were, indeed, Normal (i.e Skewness = 0 and Kurtosis = 3).

For Q, again, the skewness is not inconsistent with Normality. However, this is not the case for the kurtosis, with all of the test-statistic distributions having kurtosis greater than would be expected for Normal data.

Figures 6.1 - 6.13 show histograms of Q for the various defined options for the matching function. In these histograms, broken lines denote the expected form of the histograms if Q was, truly, Standard Normal. The simplest summary of these would be that "Sometimes the Normal approximation to Q appears to be better than others".

More specifically, away from the more extreme choices of $c = 0.05$ or 0.25 and $k = 20$ or 7.5 (7.5 to a lesser extent), the observed distribution of Q does not deviate too seriously from its Standard Normal approximation, although there are noticable differences in the distribution tails.

To assess the seriousness of the deviations from the Normal

Table 6.1 : Mean, Variance, Skewness and Kurtosis from Simulated Null Distribution of Marcus and Genizi's W Statistic

(i) Caliper Matching

Matching Function Specification	MEAN	VARIANCE	SKEWNESS	KURTOSIS
0.05 x IQR	0.500	0.010	0.021	2.996
0.25 x IQR	0.500	0.005	0.017	2.973
0.50 x IQR	0.500	0.004	0.006	3.015
0.625 x IQR	0.500	0.003	0.006	3.026
0.75 x IQR	0.499	0.003	0.006	3.024
1.50 x IQR	0.500	0.003	-0.014	2.967
2.50 x IQR	0.500	0.004	-0.014	2.958
Infinite Caliper	0.500	0.004	-0.014	2.950

(ii) Continuous Matching Function

exp(-20d / IQR)	0.500	0.007	0.014	2.968
exp(-7.5d / IQR)	0.500	0.005	0.014	2.978
exp(-5d / IQR)	0.500	0.004	0.011	2.990
exp(-2.5d / IQR)	0.500	0.003	0.009	2.996
exp(-0.5d / IQR)	0.500	0.003	-0.007	2.956

NOTE : s.e.(skewness) = $\sqrt{(6/n)}$

Compare observed values to 1.96 x s.e = 0.034

s.e.(kurtosis) = $\sqrt{(24/n)}$

Compare observed values to 3 + (1.96 x s.e.) = 3.068

(based on 20000 simulations)

Table 6.2 : Mean, Variance, Skewness and Kurtosis from Simulated Null Distribution of Marcus and Genizi's Q Statistic

(i) Caliper Matching

Matching Function Specification	MEAN	VARIANCE	SKEWNESS	KURTOSIS
0.05 x IQR	0.004	0.816	0.000	4.380
0.25 x IQR	-0.002	0.977	0.016	3.480
0.50 x IQR	-0.003	1.026	0.000	3.371
0.625 x IQR	-0.002	1.030	-0.001	3.329
0.75 x IQR	-0.001	1.032	0.000	3.273
1.50 x IQR	0.001	1.060	-0.010	3.182
2.50 x IQR	0.002	1.096	-0.012	3.301
Infinite Caliper	0.003	1.105	-0.013	3.331

(ii) Continuous Matching Function

exp(-20d / IQR)	0.003	0.896	0.007	4.013
exp(-7.5d / IQR)	0.000	0.992	0.006	3.589
exp(-5d / IQR)	-0.001	1.020	0.006	3.493
exp(-2.5d / IQR)	-0.001	1.043	0.004	3.342
exp(-0.5d / IQR)	0.002	1.071	-0.003	3.203

NOTE : s.e.(skewness) = $\sqrt{(6/n)}$

Compare observed values to 1.96 x s.e. = 0.034

s.e.(kurtosis) = $\sqrt{(24/n)}$

Compare observed values to 3 + (1.96 x s.e.) = 3.068

(based on 20000 simulations)

Figure 6.1 : Histogram of Q ; Caliper = 0.05 * IQR

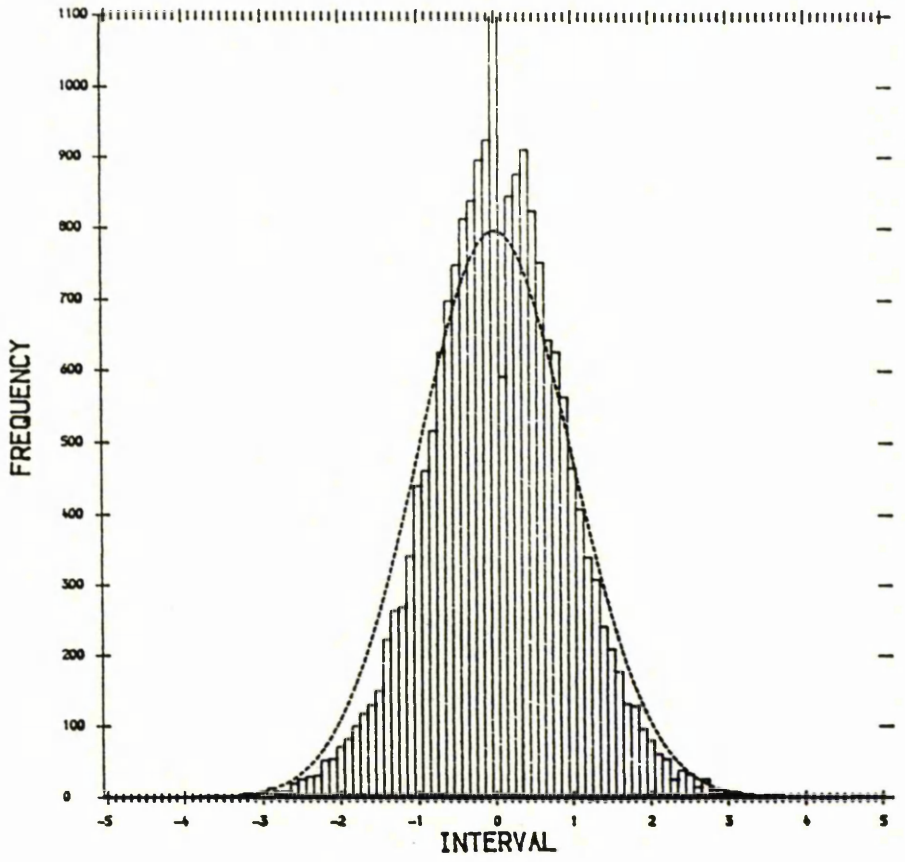


Figure 6.2 : Histogram of Q ; Caliper = 0.25 * IQR

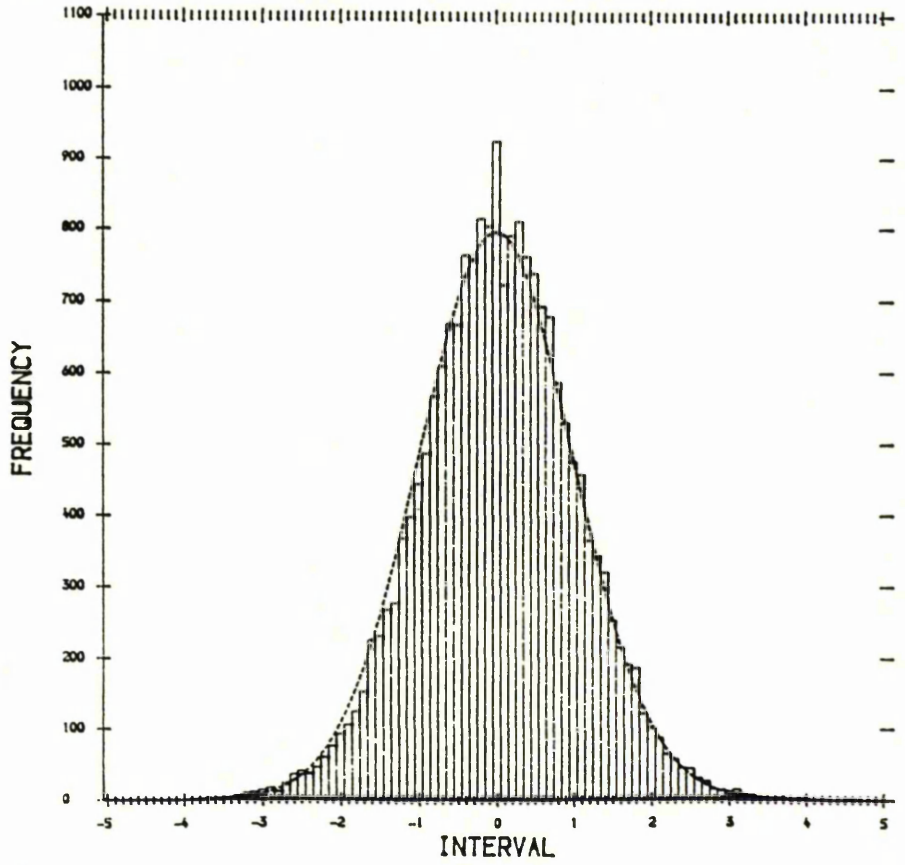


Figure 6.3 : Histogram of Q : Caliper = 0.50 * IQR

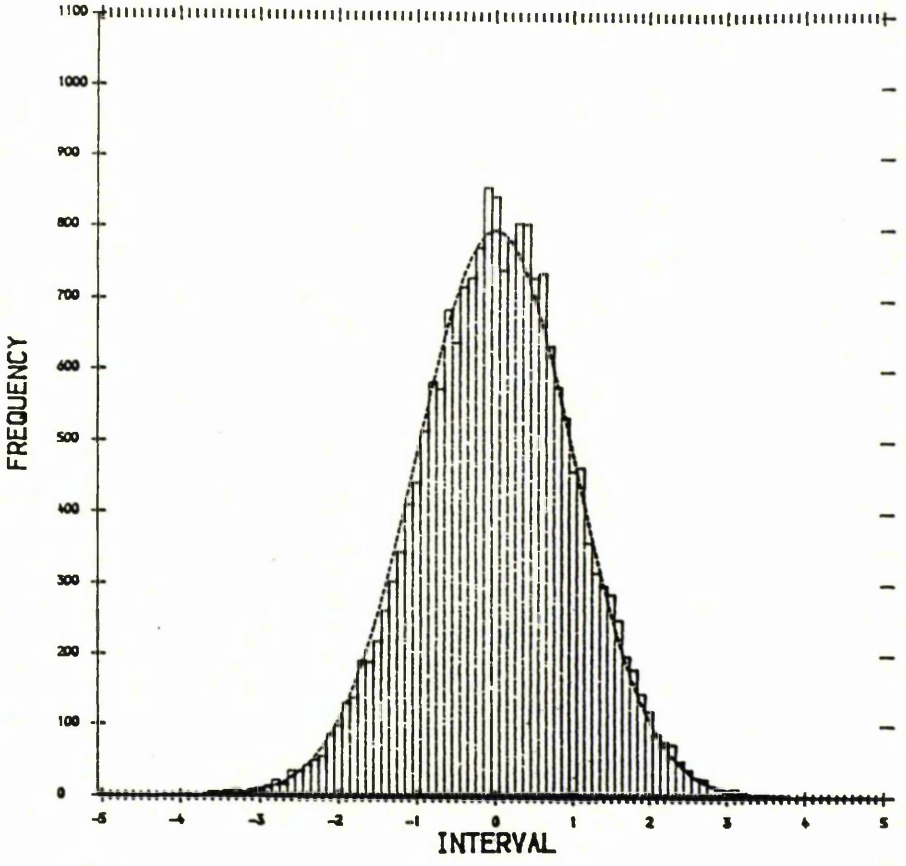


Figure 6.4 : Histogram of Q : Caliper = 0.625 * IQR

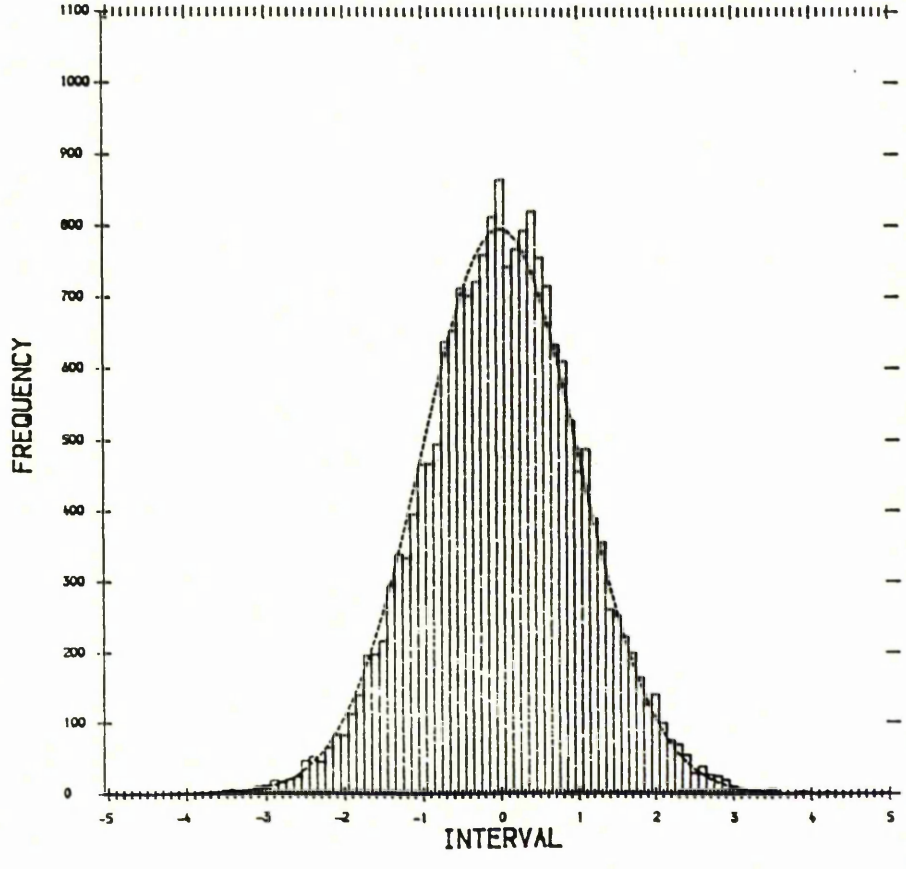


Figure 6.5 : Histogram of Q ; Caliper = 0.75 * IQR

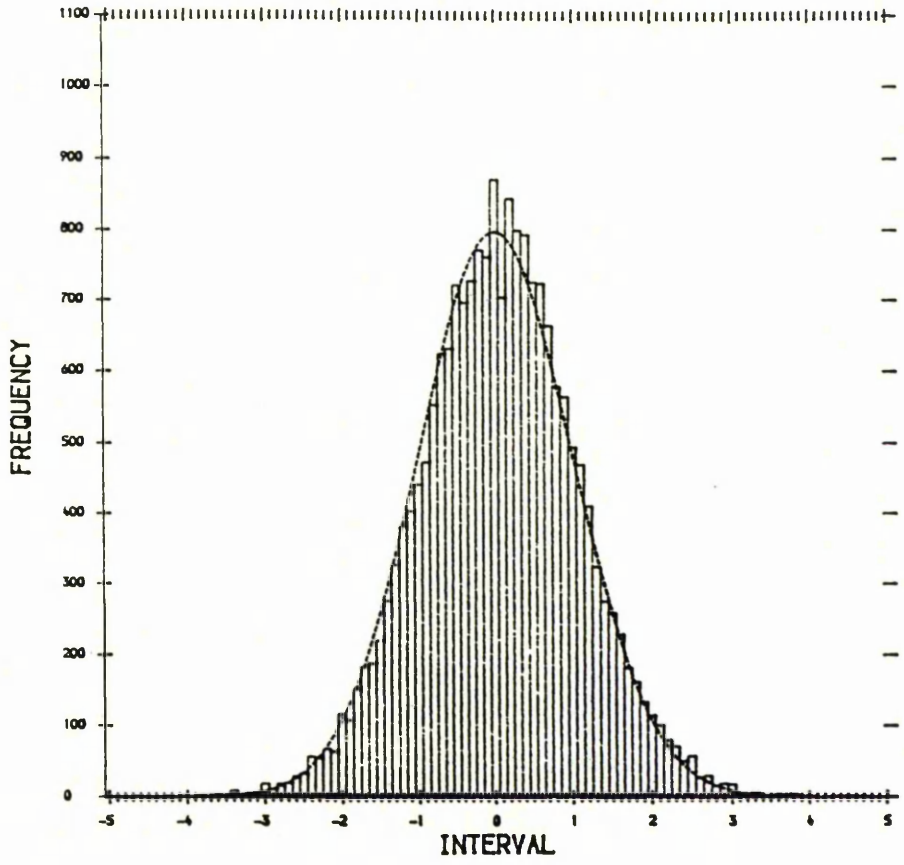


Figure 6.6 : Histogram of Q ; Caliper = 1.50 * IQR

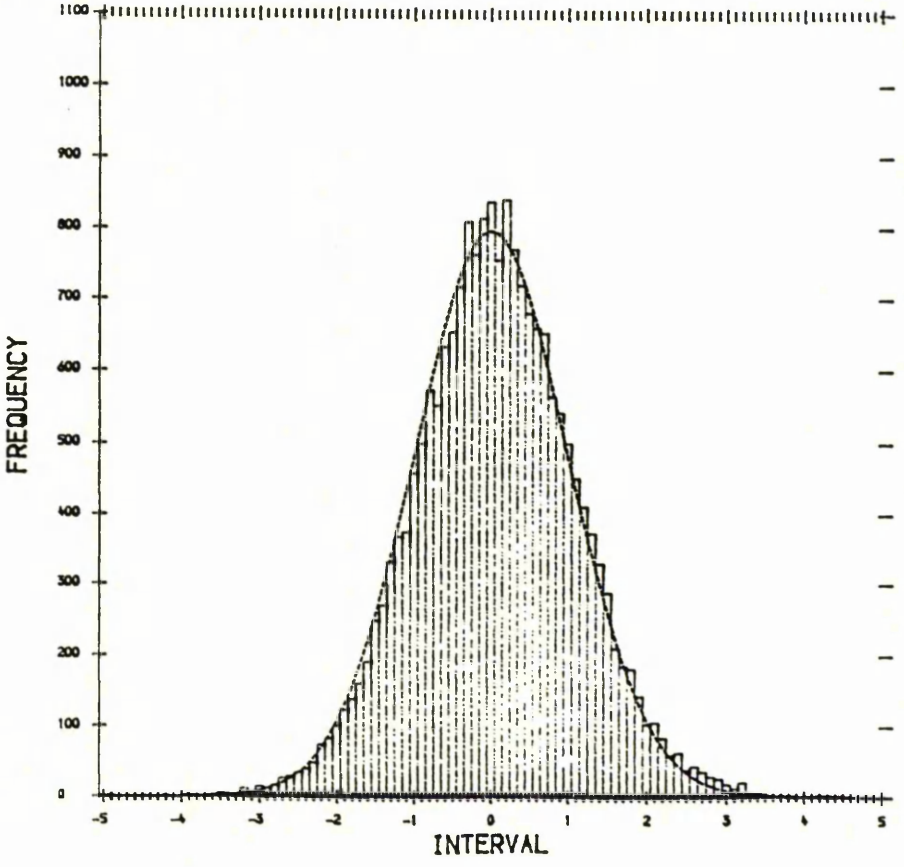


Figure 6.7 : Histogram of Q ; Caliper = 2.50 * IQR

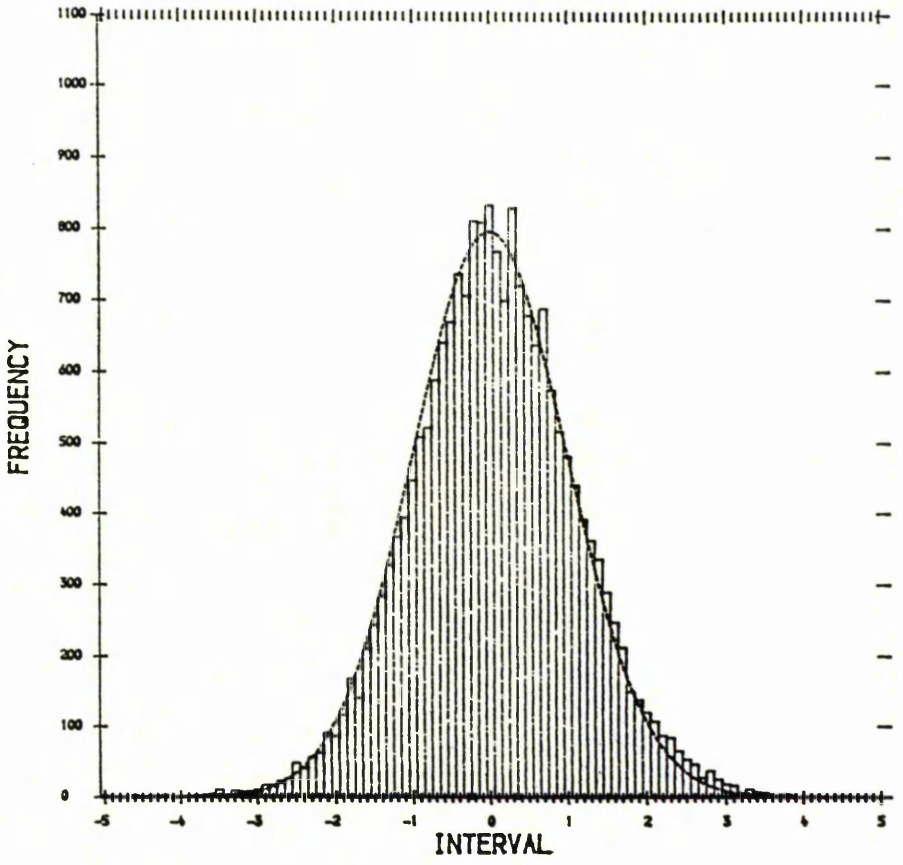


Figure 6.8 : Histogram of Q ; Infinite Caliper

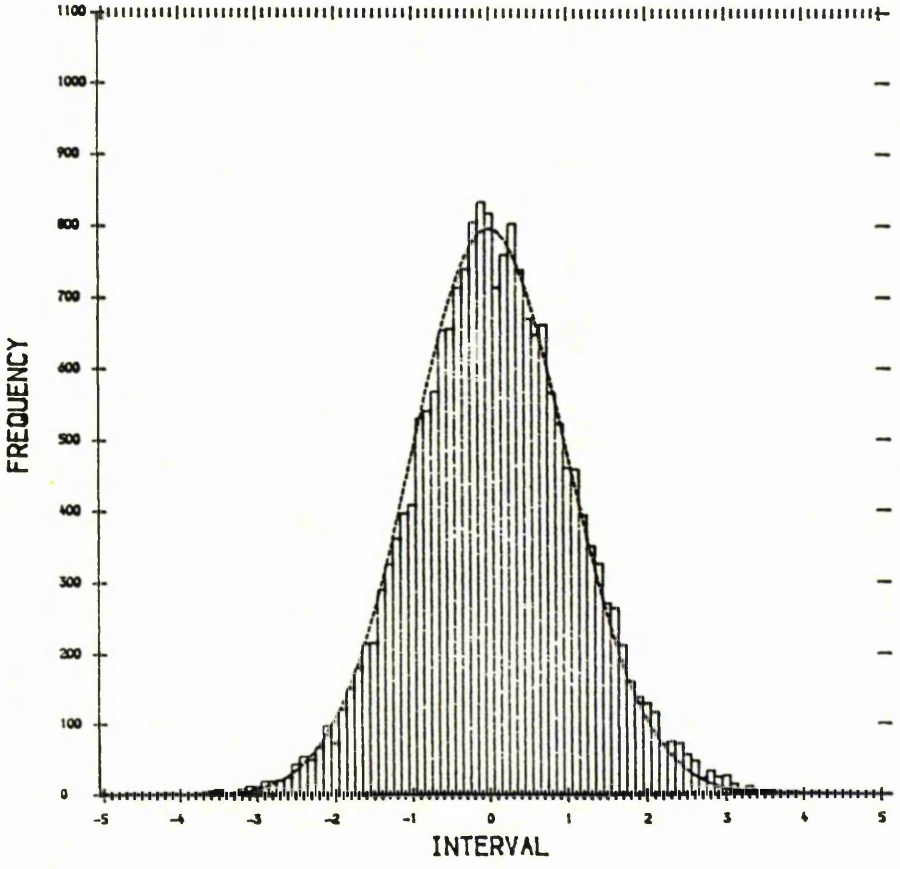


Figure 6.9 : Histogram of Q ; M.F. = exp(-20d / IQR)

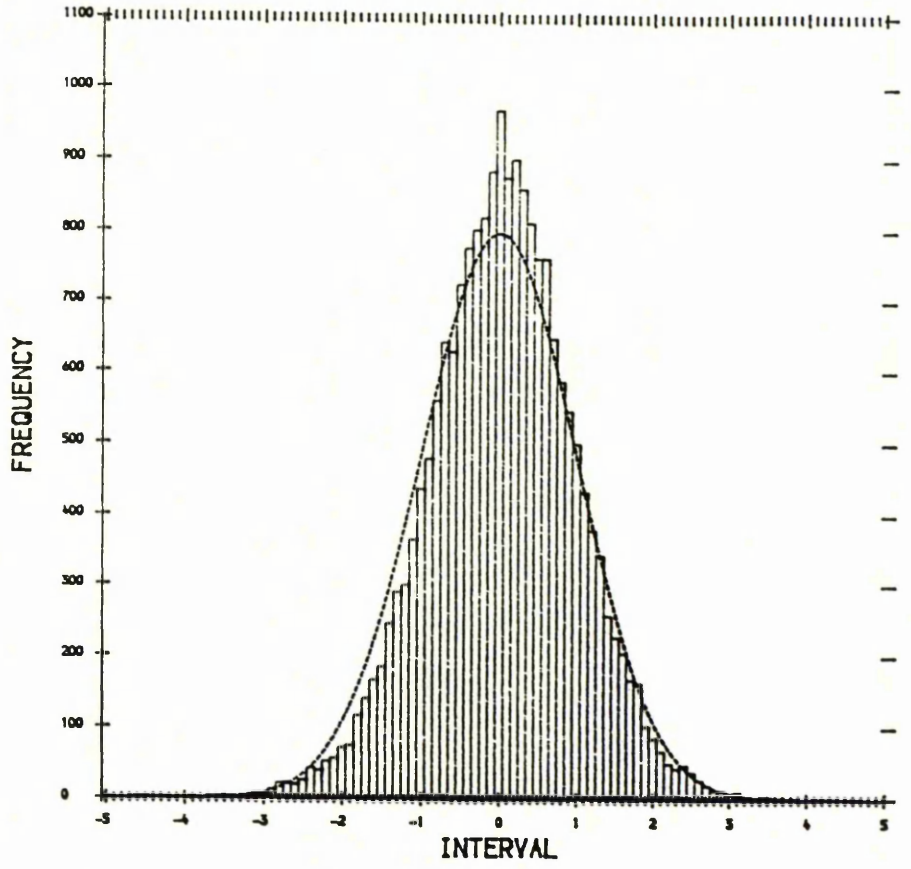


Figure 6.10 : Histogram of Q ; M.F. = exp(-7.5d / IQR)

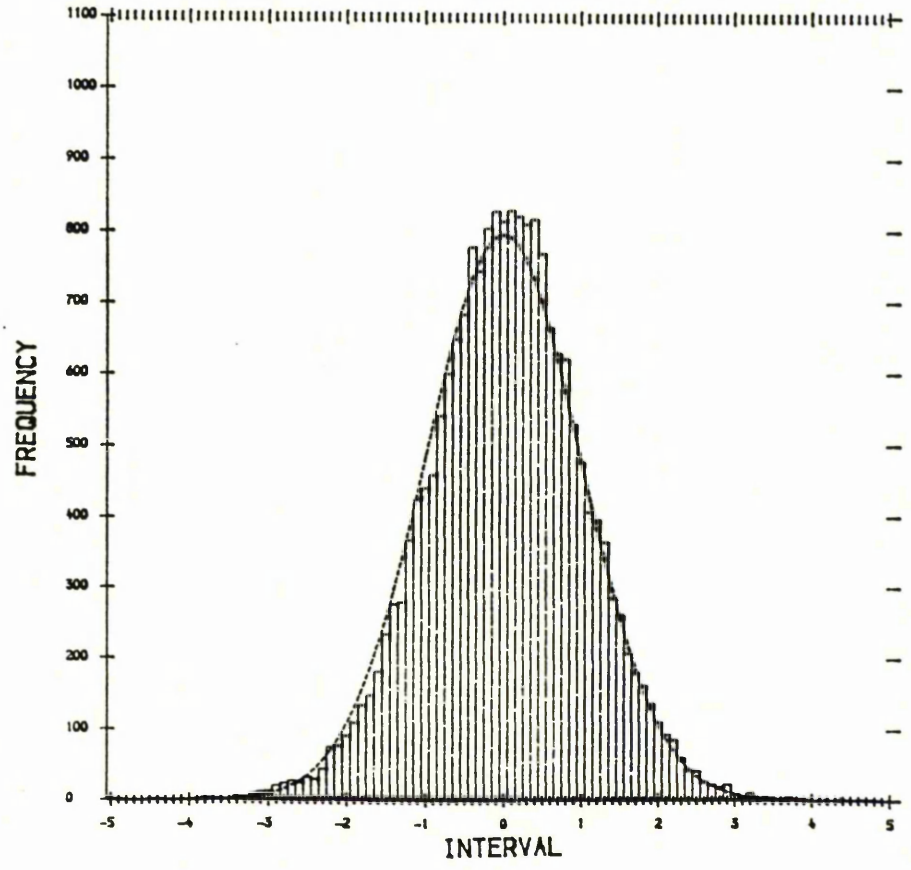


Figure 6.11 : Histogram of Q : M.F. = exp(-5d / IQR)

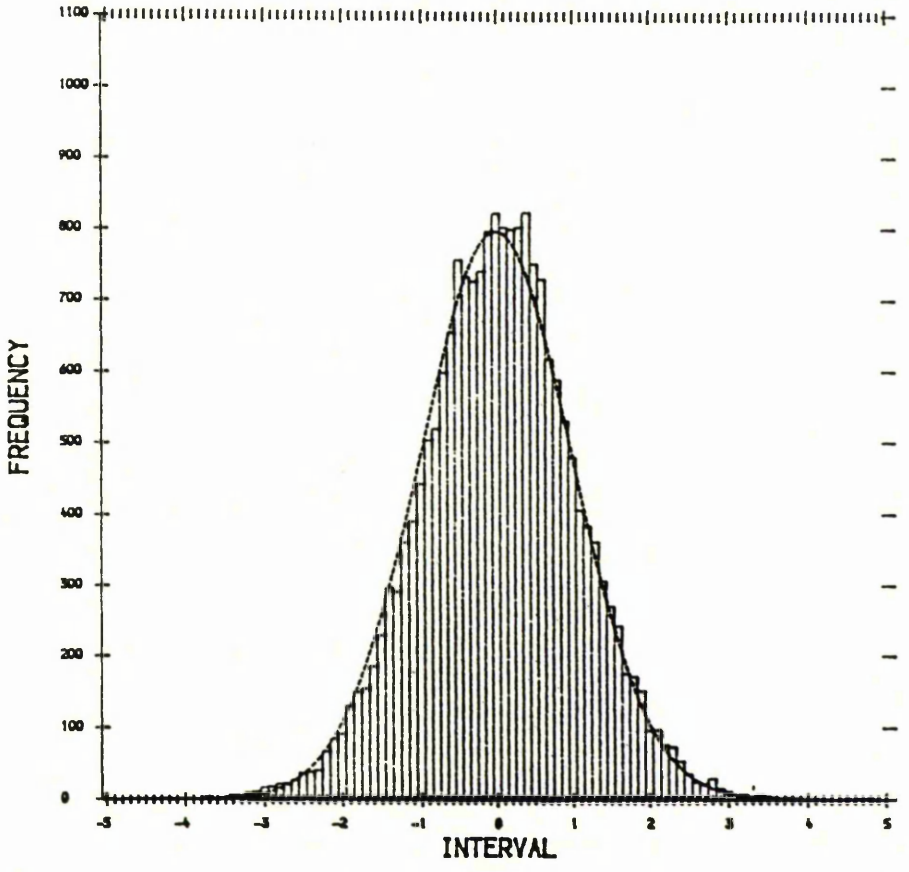


Figure 6.12 : Histogram of Q : M.F. = exp(-2.5d / IQR)

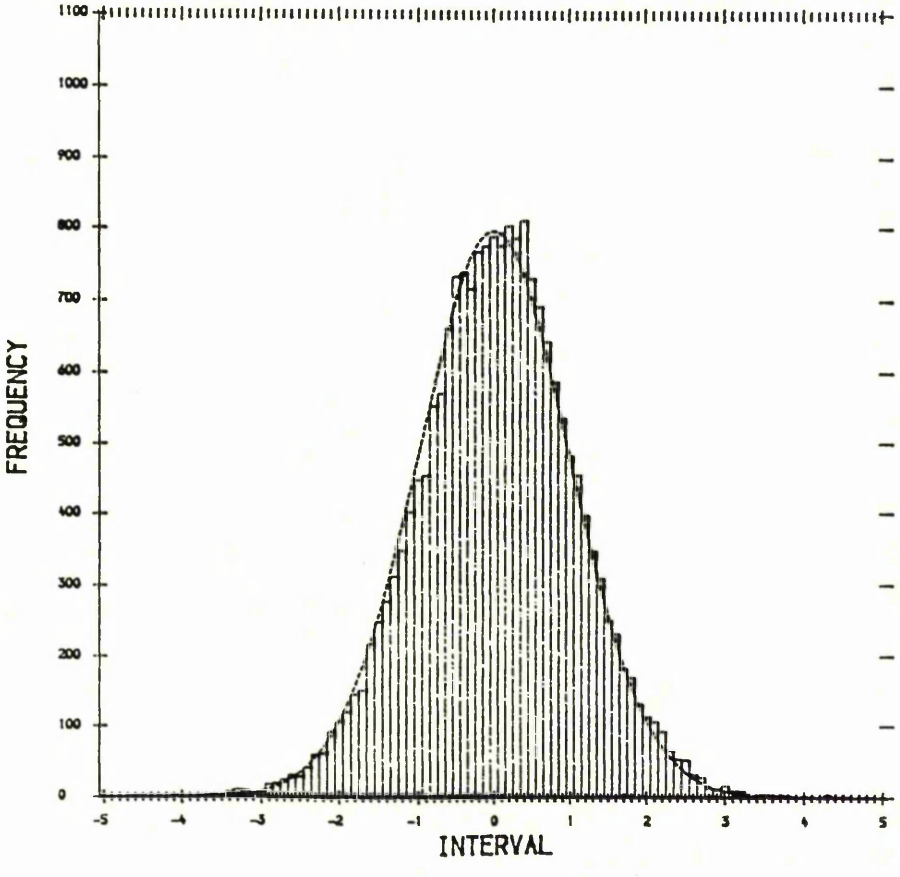
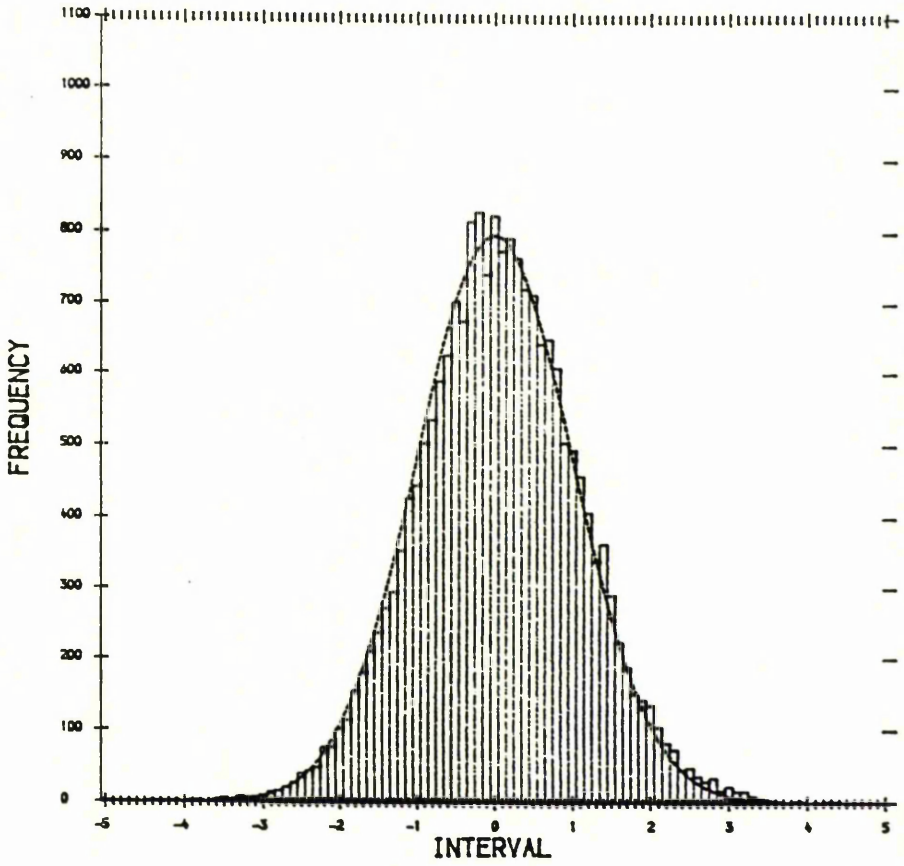


Figure 6.13 : Histogram of Q ; M.F. = exp(-0.5d / IQR)



approximations, the proportions of the observed values of Q (for each option seperately) lying above the 95th and 90th percentiles of the Standard Normal distribution were calculated. These results are shown in Table 6.3 & 6.4 . (If the N(0,1) approximation was exactly correct, then 5% of the values of Q would be expected to lie above the 95th percentile, and 10% would be expected to lie above the 90th percentile).

It can be seen that, again, at the more "extreme" choices of matching function, the approximation was fairly poor (e.g for $c = 0.05$, only 3.5% of the values of Q lay above the 95th percentile, and only 7.25% lay above the 90th percentile (instead of the nominal values of 5% and 10% respectively)) . For the less extreme choices of c or k, the proportion of tests which would be rejected using Normal critical values were similar to the nominal values.

The 95th and 90th percentiles for the simulated distributions of Q and W are shown in Tables 6.5 & 6.6 (these being the 19000th and 18000th values, respectively, in the simulated distributions).

The conclusions from Part (A) would be that, although, obviously, a Normal approximation for Marcus and Genizi's tests is not ideal in some cases, in many cases the observed deviations from that approximation were not sufficiently serious to merit disregarding the Standard Normal approximation for Q (and the advantages of simplicity that such an approximation would bring).

However, due to the inadequacies of the Normal approximation for more extreme choices of c and k, it was decided to carry forward to Part (B) the critical values derived by simulation, so that meaningful interpretation could be made of the results using these extreme choices, and so that comparison could be made of all of the test options on an equal footing.

Also, since using Q brings in the complication of estimating the standard deviation of W at each stage, it was decided to use the statistic, W, rather than its standardised version, Q.

Model Specifications for (B)

Of interest was to make power comparisons of the various Marcus and Genizi options. Carried forward from Part (A) were the

Table 6.3 : Proportion of simulations where Q exceeded the N(0,1) 90th and 95th Percentiles (20000 simulations)

(20 cases per group)

(i) Caliper Matching

Matching Function Specification	Proportion larger than 95th Percentile	Proportion larger than 90th Percentile
0.05 x IQR	3.54 %	7.25 %
0.25 x IQR	4.76 %	9.56 %
0.50 x IQR	5.12 %	10.02 %
0.625 x IQR	5.16 %	9.60 %
0.75 x IQR	5.08 %	9.62 %
1.50 x IQR	5.49 %	10.27 %
2.50 x IQR	5.75 %	10.90 %
Infinite Caliper	5.70 %	11.00 %

(ii) Continuous Matching Function

exp(-20d / IQR)	4.00 %	8.20 %
exp(-7.5d / IQR)	4.90 %	9.44 %
exp(-5d / IQR)	5.04 %	9.86 %
exp(-2.5d / IQR)	5.26 %	9.96 %
exp(-0.5d / IQR)	5.48 %	10.61 %

NOTE : If a given test's distribution was truly Standard Normal, the proportion of tests "rejected" at 5% and 10% would be expected to be within 5% $\pm$ 0.3% and 10% $\pm$ 0.4% respectively.

Table 6.4 : Proportion of simulations where Q exceeded the N(0,1) 90th and 95th Percentiles (20000 simulations)

(50 cases per group)

(i) Caliper Matching

Matching Function Specification	Proportion larger than 95th Percentile	Proportion larger than 90th Percentile
0.05 x IQR	3.38 %	7.65 %
0.25 x IQR	4.80 %	9.56 %
0.50 x IQR	4.93 %	9.78 %
0.625 x IQR	4.90 %	9.90 %
0.75 x IQR	4.99 %	9.87 %
1.50 x IQR	4.89 %	9.90 %
2.50 x IQR	4.92 %	9.82 %
Infinite Caliper	5.08 %	9.83 %

(ii) Continuous Matching Function

exp(-20d / IQR)	4.04 %	8.64 %
exp(-7.5d / IQR)	4.74 %	9.68 %
exp(-5d / IQR)	4.96 %	9.88 %
exp(-2.5d / IQR)	5.02 %	9.74 %
exp(-0.5d / IQR)	4.90 %	9.72 %

NOTE : If a given test's distribution was truly Standard Normal, the proportion of tests "rejected" at 5% and 10% would be expected to be within 5% $\pm$ 0.3% and 10% $\pm$ 0.4% respectively.

Table 6.5 : Critical Values for Marcus & Genizi Statistic, W

(i) CALIPER MATCHING

Matching Function Spec.	5 %	10 %
.05 x IQR	0.660	0.628
.25 x IQR	0.614	0.589
.50 x IQR	0.599	0.576
.625 x IQR	0.593	0.572
.75 x IQR	0.589	0.569
1.5 x IQR	0.589	0.569
2.5 x IQR	0.598	0.577
Infinite Caliper	0.600	0.579

(ii) CONTINUOUS MATCHING

exp(-20L)	0.634	0.605
exp(-7.5L)	0.612	0.587
exp(-5L)	0.603	0.581
exp(-2.5L)	0.592	0.571
exp(-0.5L)	0.590	0.570

where L = d / IQR

(Based on 20000 simulations)

Table 6.6 : Critical Values for Marcus & Genizi Statistic, Q

(i) CALIPER MATCHING

Matching Function Spec.	5 %	10 %
.05 x IQ	1.489	1.119
.25 x IQR	1.611	1.229
.50 x IQR	1.655	1.262
.625 x IQR	1.669	1.274
.75 x IQR	1.671	1.280
1.5 x IQR	1.673	1.288
2.5 x IQR	1.702	1.311
Infinite Caliper	1.704	1.322

(ii) CONTINUOUS MATCHING

exp(-20L)	1.533	1.171
exp(-7.5L)	1.617	1.234
exp(-5L)	1.651	1.249
exp(-2.5L)	1.677	1.273
exp(-0.5L)	1.693	1.294

where L = d / IQR

(Based on 20000 simulations)

simulated critical values for test statistic, W. Although the general form of the simulation model was as before, the intercepts for the groups were separated by known amounts (measured in units of σ_y). The intercepts were equally spaced, with $\alpha_1 = -15$ and $\alpha_1 \leq \alpha_2 \leq \alpha_3$.

Starting with $\alpha_1 = \alpha_2 = \alpha_3$, simultaneously $(\alpha_2 - \alpha_1)$ was increased in steps of $0.125\sigma_y$ and $(\alpha_3 - \alpha_1)$ was increased in steps of $0.25\sigma_y$.

i.e

Step 0	$\alpha_1 = -15$	$\alpha_2 = -15$	$\alpha_3 = -15$
Step 1	$\alpha_1 = -15$	$\alpha_2 = \alpha_1 + 0.125\sigma_y$	$\alpha_3 = \alpha_1 + 0.25\sigma_y$
Step 2	$\alpha_1 = -15$	$\alpha_2 = \alpha_1 + 0.25\sigma_y$	$\alpha_3 = \alpha_1 + 0.50\sigma_y$
⋮			
Step 8	$\alpha_1 = -15$	$\alpha_2 = \alpha_1 + \sigma_y$	$\alpha_3 = \alpha_1 + 2\sigma_y$

For every choice of α_1, α_2, α_3, 7500 simulations were performed, and the proportion of times that each form of test statistic was rejected was calculated (where test statistics would be "rejected" if they were greater than their own specific critical value). The results are shown in Table 6.7. The conclusions from Part (B) are as shown below :

The Results for (B)

(i) Away from the more extreme choices of c or k, the tests are fairly insensitive to the choice of matching function made.

(ii) For the caliper matching approach, a reasonable choice of c was 0.5 (although this choice would not be too critical).

(iii) For the approach using a continuous matching function, k = 5 would seem to be the most suitable choice, although, again, this choice would not be too critical.

(iv) It would appear to be more serious to choose a value of k too small than to choose too large a value - the power of test falls away more rapidly as k decreases than it does as k increases away from its "optimal" value (see the plots of Power vs k in Figures 6.14 & 6.15).

Due to its relatively favourable performance and the intuitive desirability of having a continuous matching function, it was

Table 6.7 : Comparison of the Marcus & Genizi Options - The Proportion of Tests Rejected at 5 %
(7500 simulations)

(A) Caliper Matching

	ALPHA 3 - ALPHA 1								
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
(i)	4.9	27.7	67.1	92.3	98.9	99.9	100.0	100.0	100.0
(ii)	5.0	42.0	89.0	99.7	100.0	100.0	100.0	100.0	100.0
(iii)	5.3	44.4	91.1	99.8	100.0	100.0	100.0	100.0	100.0
(iv)	5.1	44.1	90.5	99.8	100.0	100.0	100.0	100.0	100.0
(v)	5.1	42.0	89.3	99.8	100.0	100.0	100.0	100.0	100.0
(vi)	5.3	27.2	65.4	91.9	98.9	99.9	100.0	100.0	100.0
(vii)	5.2	20.5	48.9	78.1	94.3	98.9	99.9	100.0	100.0
(viii)	5.2	19.3	45.8	74.7	92.6	98.4	99.9	100.0	100.0

(B) Continuous Matching Function

(ix)	5.0	35.6	80.0	98.0	99.9	100.0	100.0	100.0	100.0
(x)	5.3	42.5	89.1	99.7	100.0	100.0	100.0	100.0	100.0
(xi)	5.4	44.2	90.8	99.8	100.0	100.0	100.0	100.0	100.0
(xii)	5.1	43.1	89.9	99.8	100.0	100.0	100.0	100.0	100.0
(xiii)	5.2	27.8	66.4	92.4	99.1	99.9	100.0	100.0	100.0

The Matching Function Specifications

(i) 0.05 x IQR
(ii) 0.25 x IQR
(iii) 0.50 x IQR
(iv) 0.625 x IQR
(v) 0.75 x IQR
(vi) 1.5 x IQR
(vii) 2.5 x IQR
(viii) Infinite Caliper
(ix) exp(-20d / IQR)
(x) exp(-7.5d / IQR)
(xi) exp(-5.0d / IQR)
(xii) exp(-2.5d / IQR)
(xiii) exp(-0.5d/ IQR)

Alpha 3 - Alpha 1 (s.d. units)

(a) 0.00 (b) 0.25 (c) 0.50 (d) 0.75 (e) 1.00
(f) 1.25 (g) 1.50 (h) 1.75 (i) 2.00

Figure 6.14 : Power vs k $\alpha_3-\alpha_1 = 0.25\sigma_y$

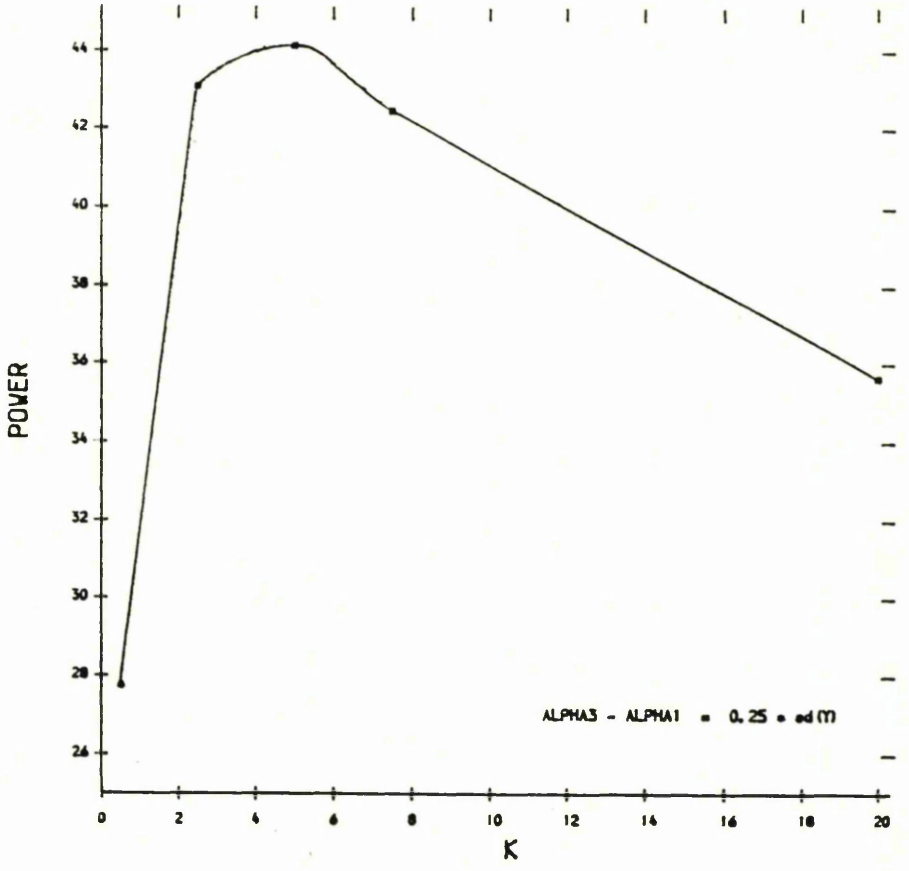
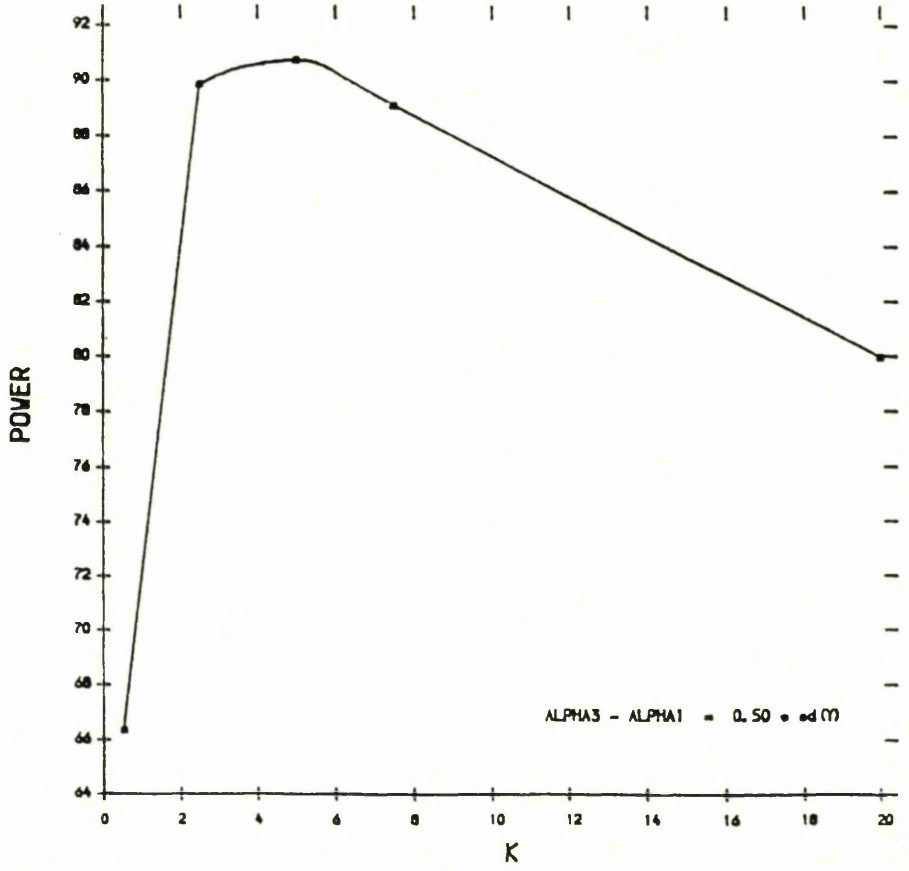


Figure 6.15 : Power vs k $\alpha_3-\alpha_1 = 0.50\sigma_y$



decided to represent Marcus and Genizi's method (Test 8 in the full testing program) by the test using a continuous matching function of form :

$$M(X_{ij},X_{i'j'}) = \exp\left[- \frac{5\ D(X_{ij},X_{i'j'})}{IQR} \right]$$

It must, of course, be borne in mind that the "ideal" choice of k would be dependent on the sample size, the covariate variability, etc., as was discussed previously in Section 6.2.

6.3 : The Main Study

6.3.1 : Aims and Objectives

Recall that of main interest was to compare the eight tests as defined in Section 5.2 earlier, under several different conditions, namely :

(1) Changing the magnitude of group spacings
(2) Changing the relative group spacings
(3) Varying the magnitude of correlation between the response and covariate variables.
(4) Varying the error distributions

6.3.2 : The Procedures Followed

Procedure for objectives (1) to (3)

For $\rho^2 = 0,\ .2,\ .4,\ .6,\ .8,\ .99$ and using a basic regression model as described earlier, with Normal errors :

(a) The intercepts were varied as in Part (B) of the previous section (i.e equal spacing of the intercepts, measured in units of σ_y).
(b) The intercepts were varied such that $\alpha_1 = \alpha_2$, and α_3, the intercept for Group 3, increased in steps of $0.25\sigma_y$.

The results of this section are shown in Tables 6.8 - 6.19 , Figures 6.16-6.27 , and are summarised within Section 6.3.5.

Table 6.8 : Normal Errors, $\rho^2 = 0$, Equal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.1	9.4	26.6	53.4	79.1	94.2	99.2	99.9	100.0
TEST 2	5.0	9.3	25.8	52.6	78.2	93.8	99.9	99.9	100.0
TEST 3	4.8	9.1	25.1	51.3	76.7	92.7	98.7	99.8	100.0
TEST 4	5.1	9.4	25.5	51.3	76.5	92.6	98.6	99.8	100.0
TEST 5	4.8	17.8	44.3	72.6	91.5	98.3	99.9	100.0	100.0
TEST 6	5.2	18.9	45.7	73.3	91.5	98.1	99.9	100.0	100.0
TEST 7	4.9	18.1	45.4	74.2	91.9	98.2	99.9	100.0	100.0
TEST 8	5.1	16.4	39.6	66.1	85.5	95.6	99.2	99.9	100.0

Table 6.9 : Normal Errors, $\rho^2 = 0$, Unequal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.1	11.1	33.7	67.5	90.3	98.5	99.9	100.0	100.0
TEST 2	5.0	10.8	33.1	66.3	89.5	98.3	99.8	100.0	100.0
TEST 3	4.8	10.6	31.2	64.3	88.2	97.7	99.7	100.0	100.0
TEST 4	5.1	10.8	31.9	64.0	88.2	97.8	99.7	100.0	100.0
TEST 5	4.8	18.8	49.2	80.5	96.5	99.5	100.0	100.0	100.0
TEST 6	5.2	19.6	49.9	81.2	96.2	99.5	100.0	100.0	100.0
TEST 7	4.9	18.0	43.5	72.6	91.0	97.9	99.7	100.0	100.0
TEST 8	5.1	16.5	38.9	65.1	84.6	95.3	98.7	99.7	99.9

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Table 6.10 : Normal Errors, $\rho^2 = 0.20$, Equal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.9	10.0	26.6	52.8	79.6	94.2	99.1	99.9	100.0
TEST 2	5.2	11.1	30.8	62.2	86.8	97.7	99.7	99.9	100.0
TEST 3	4.8	9.2	25.1	49.9	76.6	92.5	98.7	99.9	100.0
TEST 4	5.1	10.5	29.3	58.9	84.1	96.7	99.5	100.0	100.0
TEST 5	5.1	18.5	43.5	72.0	91.5	98.3	99.8	100.0	100.0
TEST 6	5.5	22.1	51.8	80.5	95.3	99.6	100.0	100.0	100.0
TEST 7	5.2	18.6	45.4	73.3	91.8	98.5	99.8	100.0	100.0
TEST 8	5.2	19.7	44.8	73.2	90.9	98.3	99.8	100.0	100.0

Table 6.11 : Normal Errors, $\rho^2 = 0.20$, Unequal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.9	11.9	33.1	66.0	89.7	98.5	99.9	100.0	100.0
TEST 2	5.2	13.2	40.4	75.9	94.7	99.7	99.9	100.0	100.0
TEST 3	4.8	11.5	31.7	63.2	87.9	98.0	99.7	100.0	100.0
TEST 4	5.1	12.7	38.3	72.4	93.4	99.3	99.9	100.0	100.0
TEST 5	5.1	19.8	49.0	79.6	95.5	99.6	100.0	100.0	100.0
TEST 6	5.5	23.6	57.6	87.9	98.3	99.9	100.0	100.0	100.0
TEST 7	5.2	19.0	44.4	72.6	90.7	98.0	99.6	100.0	100.0
TEST 8	5.2	19.5	44.5	72.3	90.6	97.8	99.6	99.9	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a)	0.0	(b)	0.25	(c)	0.50
(d)	0.75	(e)	1.00	(f)	1.25
(g)	1.50	(h)	1.75	(i)	2.00

Table 6.12 : Normal Errors, $\rho^2 = 0.40$, Equal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.5	9.8	26.9	53.5	79.5	94.1	99.0	99.9	100.0
TEST 2	4.9	12.5	40.0	75.9	95.3	99.6	100.0	100.0	100.0
TEST 3	4.7	9.3	24.7	50.7	77.3	92.6	98.6	99.9	100.0
TEST 4	4.7	12.3	36.9	71.3	93.4	99.0	100.0	100.0	100.0
TEST 5	4.9	18.5	43.4	72.0	91.3	98.2	99.8	100.0	100.0
TEST 6	5.0	24.6	61.9	89.8	98.9	99.9	100.0	100.0	100.0
TEST 7	5.0	18.9	44.2	73.2	92.3	98.5	99.9	100.0	100.0
TEST 8	5.3	22.1	54.1	83.1	96.6	99.9	100.0	100.0	100.0

Table 6.13 : Normal Errors, $\rho^2 = 0.40$, Unequal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.5	10.9	33.7	67.2	89.4	98.8	99.8	100.0	100.0
TEST 2	4.9	16.0	51.7	87.5	98.9	100.0	100.0	100.0	100.0
TEST 3	4.7	10.5	31.2	63.8	87.5	98.3	99.8	100.0	100.0
TEST 4	4.7	15.1	47.5	84.2	97.7	99.9	100.0	100.0	100.0
TEST 5	4.9	19.5	49.3	80.9	95.6	99.7	100.0	100.0	100.0
TEST 6	5.0	27.5	68.8	94.9	99.8	100.0	100.0	100.0	100.0
TEST 7	5.0	19.0	44.8	73.1	90.8	98.3	99.6	100.0	100.0
TEST 8	5.3	22.8	53.3	82.6	95.7	99.3	99.9	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Table 6.14 : Normal Errors, $\rho^2 = 0.60$, Equal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.1	10.2	25.6	53.1	79.7	94.5	99.1	99.9	100.0
TEST 2	5.2	17.5	56.2	91.5	99.3	100.0	100.0	100.0	100.0
TEST 3	4.8	9.5	24.1	50.6	76.9	93.4	98.7	99.9	100.0
TEST 4	5.4	15.7	50.2	86.7	98.3	100.0	100.0	100.0	100.0
TEST 5	4.7	18.1	43.4	72.9	91.4	98.3	99.9	100.0	100.0
TEST 6	5.4	32.9	76.0	97.4	99.9	100.0	100.0	100.0	100.0
TEST 7	4.6	18.8	44.3	74.1	91.8	98.5	99.9	100.0	100.0
TEST 8	5.1	28.2	67.3	93.9	99.4	100.0	100.0	100.0	100.0

Table 6.15 : Normal Errors, $\rho^2 = 0.60$, Unequal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.1	10.9	35.1	66.4	90.1	98.5	99.8	100.0	100.0
TEST 2	5.2	21.9	70.8	97.1	99.9	100.0	100.0	100.0	100.0
TEST 3	4.8	10.7	32.5	63.7	88.1	97.8	99.7	100.0	100.0
TEST 4	5.4	19.6	64.3	94.6	99.7	100.0	100.0	100.0	100.0
TEST 5	4.7	19.2	50.4	80.2	95.9	99.5	100.0	100.0	100.0
TEST 6	5.4	35.6	84.1	99.1	100.0	100.0	100.0	100.0	100.0
TEST 7	4.6	18.4	45.5	73.1	91.3	98.1	99.7	100.0	100.0
TEST 8	5.1	28.7	67.5	92.4	99.2	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Table 6.16 : Normal Errors, $\rho^2 = 0.80$, Equal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.0	9.6	27.0	53.8	79.8	94.5	99.0	99.9	100.0
TEST 2	4.9	31.8	86.7	99.8	100.0	100.0	100.0	100.0	100.0
TEST 3	4.7	9.5	25.2	51.3	77.5	93.3	98.6	99.8	100.0
TEST 4	4.7	25.9	79.2	98.9	100.0	100.0	100.0	100.0	100.0
TEST 5	4.6	18.1	44.9	72.9	91.6	98.2	99.8	100.0	100.0
TEST 6	5.3	51.9	95.5	100.0	100.0	100.0	100.0	100.0	100.0
TEST 7	4.6	18.3	45.4	74.0	92.1	98.6	99.9	100.0	100.0
TEST 8	5.1	44.2	89.9	99.6	100.0	100.0	100.0	100.0	100.0

Table 6.17 : Normal Errors, $\rho^2 = 0.80$, Unequal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.0	12.0	35.0	66.7	90.3	98.4	99.9	100.0	100.0
TEST 2	4.9	40.8	95.1	100.0	100.0	100.0	100.0	100.0	100.0
TEST 3	4.7	11.2	32.7	63.5	88.4	97.9	99.9	100.0	100.0
TEST 4	4.7	34.9	90.2	99.8	100.0	100.0	100.0	100.0	100.0
TEST 5	4.6	19.9	50.1	80.3	95.9	99.5	100.0	100.0	100.0
TEST 6	5.3	58.9	98.6	100.0	100.0	100.0	100.0	100.0	100.0
TEST 7	4.6	19.3	45.4	73.2	91.1	98.2	99.8	100.0	100.0
TEST 8	5.1	44.1	89.7	99.4	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Table 6.18 : Normal Errors, $\rho^2 = 0.99$, Equal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.9	9.9	26.1	53.7	80.4	94.7	99.2	99.9	100.0
TEST 2	5.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
TEST 3	4.8	9.1	25.2	51.5	78.2	93.1	98.9	99.9	100.0
TEST 4	5.3	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
TEST 5	4.8	17.9	44.3	73.1	92.2	98.5	99.9	100.0	100.0
TEST 6	5.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
TEST 7	4.7	18.8	45.7	74.4	92.8	98.6	99.9	100.0	100.0
TEST 8	3.2	99.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 6.19 : Normal Errors, $\rho^2 = 0.99$, Unequal Spacing

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.9	11.7	34.1	67.6	89.5	98.3	99.9	100.0	100.0
TEST 2	5.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
TEST 3	4.8	10.8	31.5	64.5	87.9	97.9	99.8	100.0	100.0
TEST 4	5.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
TEST 5	4.8	19.2	49.9	80.9	95.6	99.6	100.0	100.0	100.0
TEST 6	5.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
TEST 7	4.7	18.1	44.6	73.0	90.9	98.3	99.7	100.0	100.0
TEST 8	3.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Figure 6.16 : Normal Errors, $\rho^2 = 0$, Equal Spacing

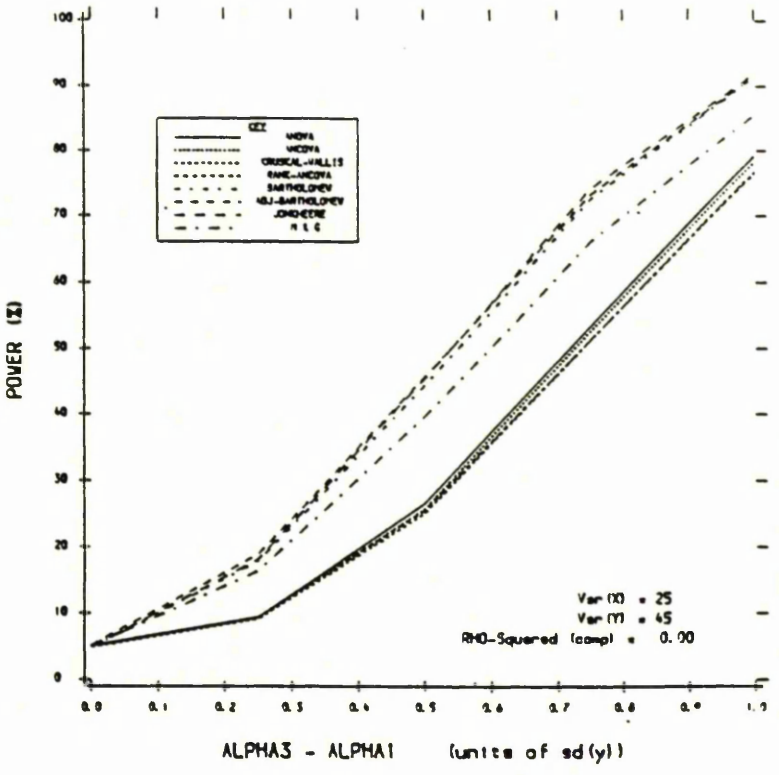


Figure 6.17 : Normal Errors, $\rho^2 = 0$, Unequal Spacing

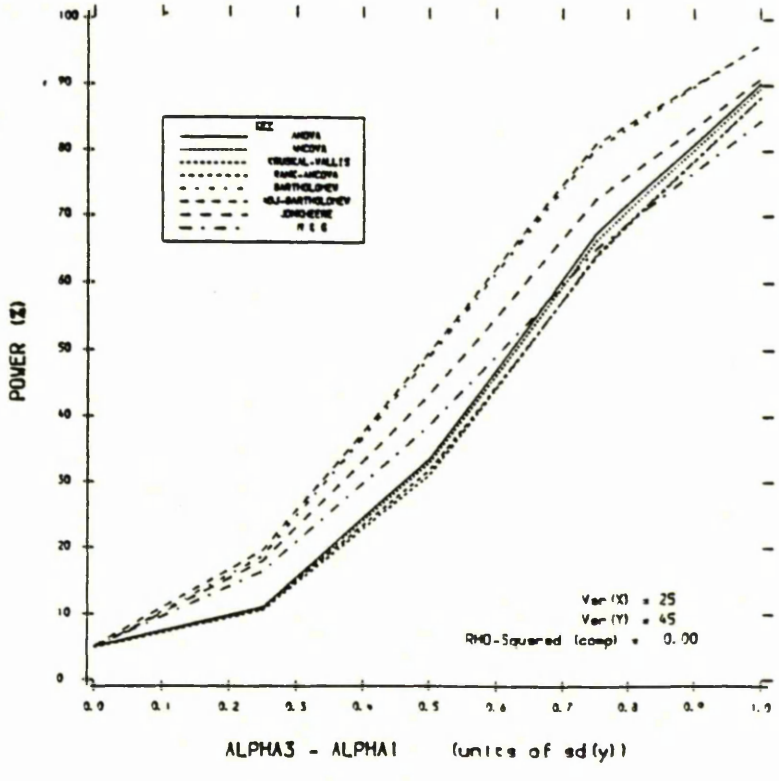


Figure 6.18 : Normal Errors, ρ^2 = 0.20, Equal Spacing

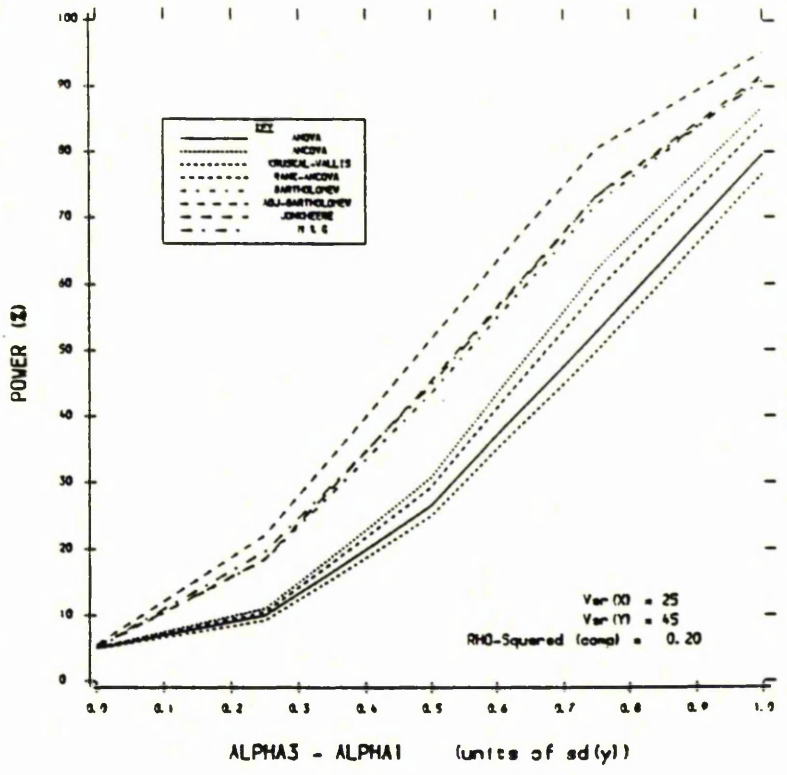


Figure 6.19 : Normal Errors, ρ^2 = 0.20, Unequal Spacing

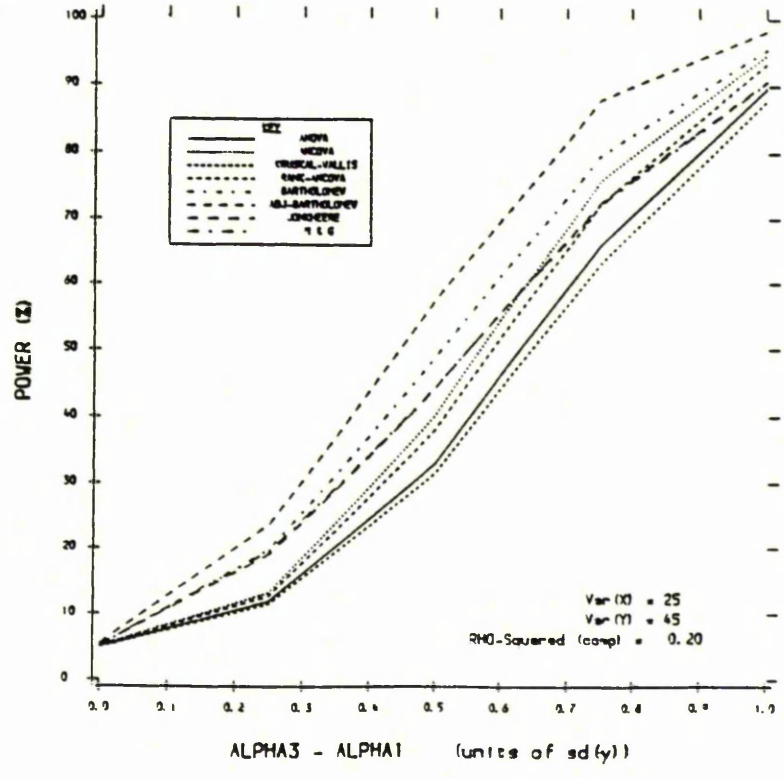


Figure 6.20 : Normal Errors, ρ^2 = 0.40, Equal Spacing

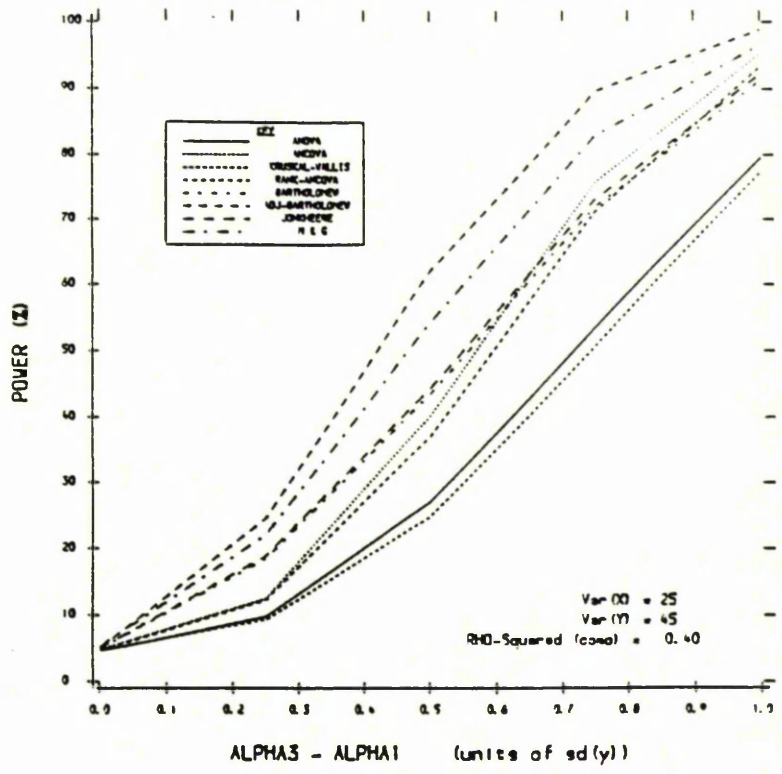


Figure 6.21 : Normal Errors, ρ^2 = 0.40, Unequal Spacing

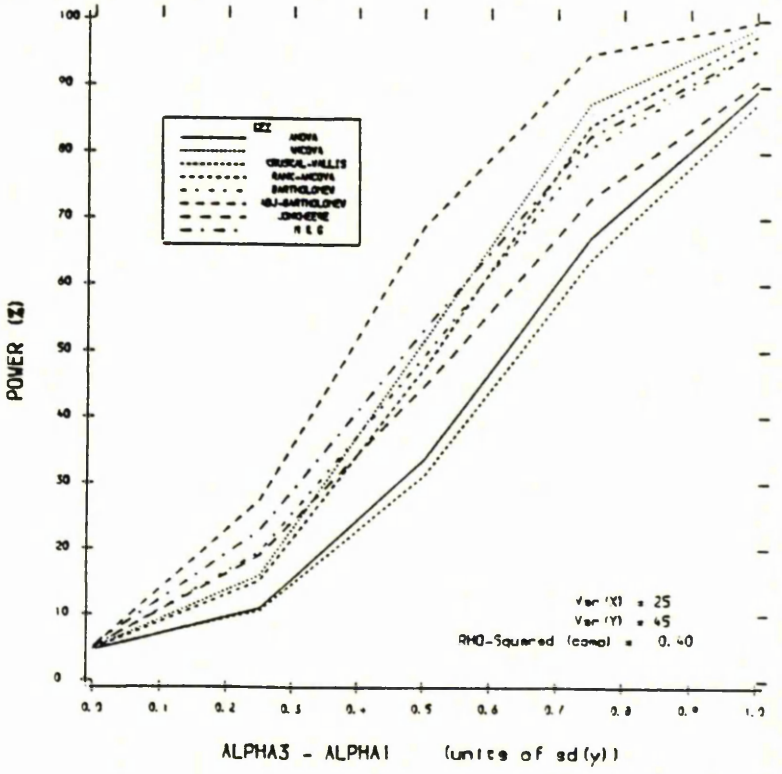


Figure 6.22 : Normal Errors, $\rho^2 = 0.60$, Equal Spacing

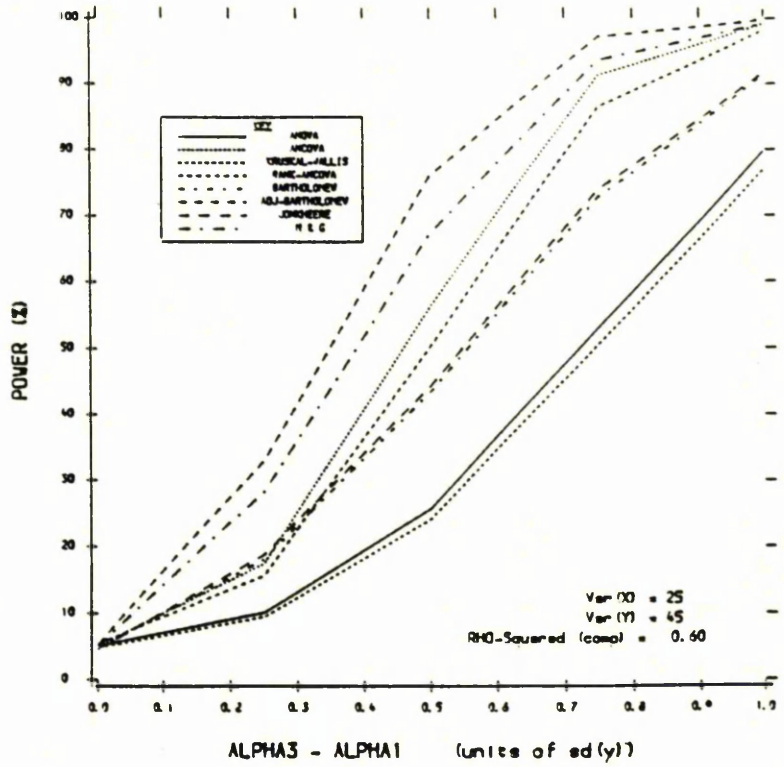


Figure 6.23 : Normal Errors, $\rho^2 = 0.60$, Unequal Spacing

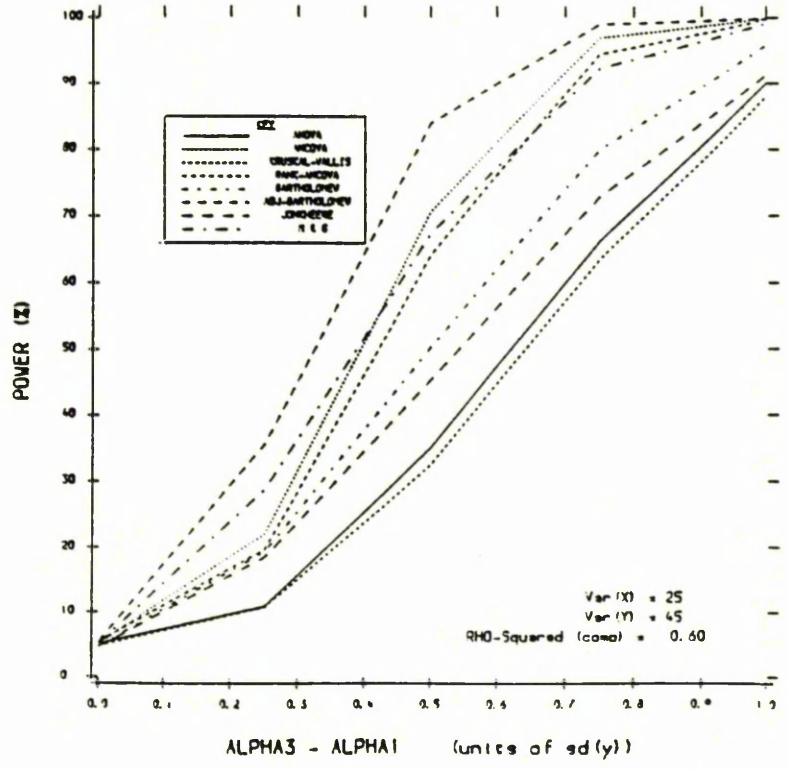


Figure 6.24 : Normal Errors, ρ^2 = 0.80, Equal Spacing

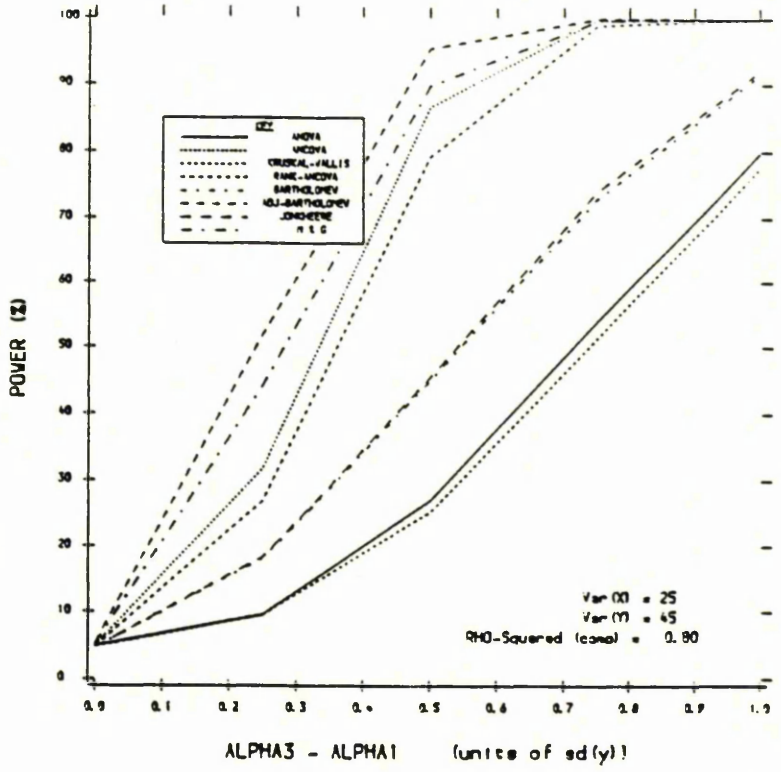


Figure 6.25 : Normal Errors, ρ^2 = 0.80, Unequal Spacing

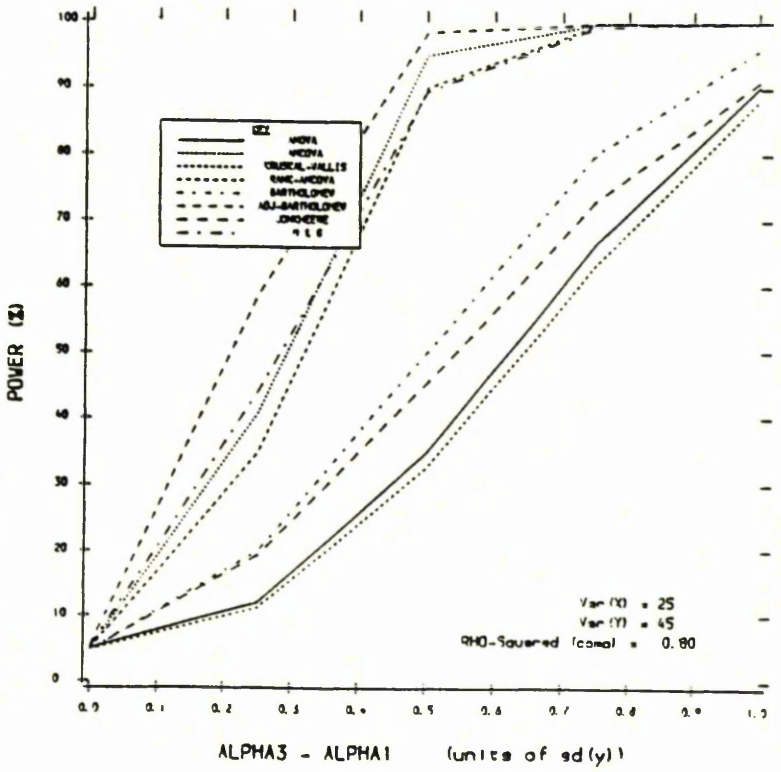


Figure 6.26 : Normal Errors, $\rho^2 = 0.99$, Equal Spacing

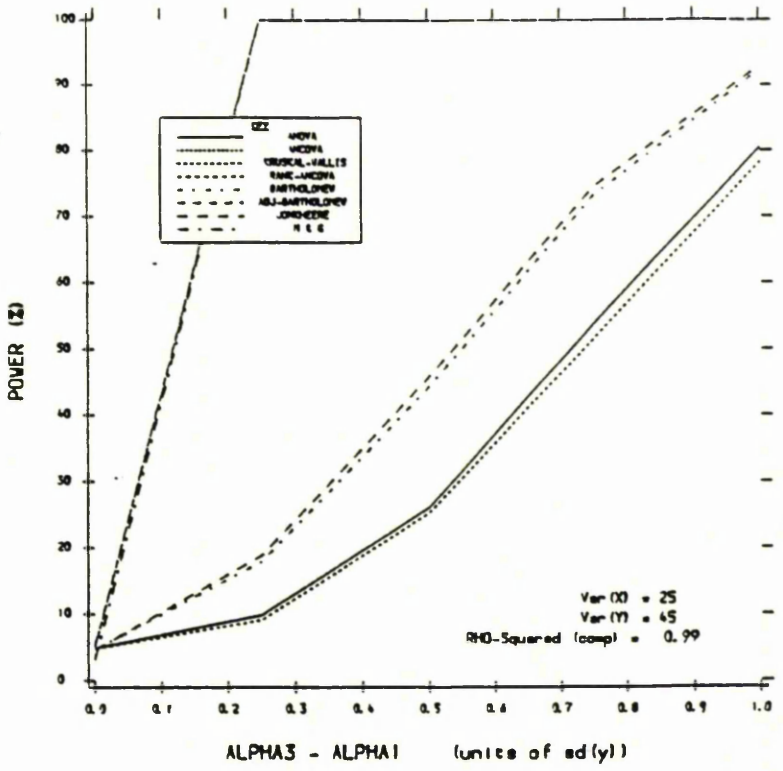
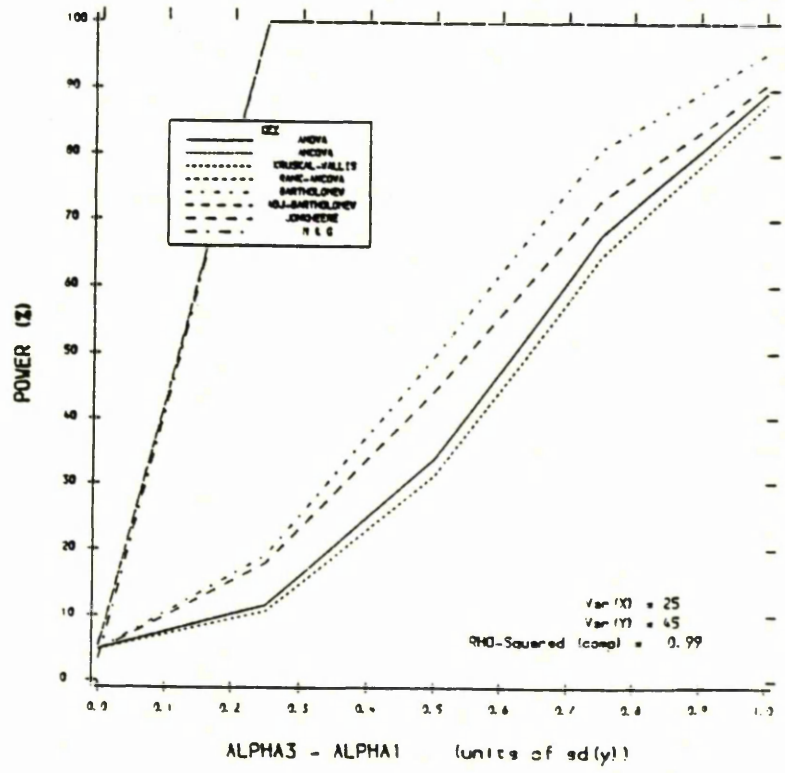


Figure 6.27 : Normal Errors, $\rho^2 = 0.99$, Unequal Spacing



Procedure for objective (4)

The work on varying the error distribution was carried out in several phases. Letting β represent the skewness of the error distribution, and γ represent its kurtosis, the phases were :

(a) Fixing $\beta = 0$ (as in the Normal distribution) while varying γ

(b) Fixing $\gamma = 3$ (as in the Normal distribution) while varying β

(c) Assessing the effect of reversing the skewness of the error distribution

(d) Varying both β and γ together

(a)-(c) were performed using mixture distributions composed of two Normal components, while (d) was performed using various Lognormal distributions.

The general procedure followed consisted of simulating errors from a distribution where β and γ were as required (but this not necessarily being the case for the mean and variance) . The simulated values were then scaled to produce values with the required mean and variance. The scaling consisted of subtracting the true mean for the simulation distribution from each simulated value, then multiplying by a factor of

$$\frac{\text{(Required Standard Deviation)}}{\text{(Standard Deviation of the Simulation Distribution)}}$$

The required standard deviation depended on the Normal-error simulations to which the results were to be compared. For example, if the aim was to compare the simulation results to those from the Normal-error simulations with $\rho^2 = 0.8$, then the simulated non-Normal errors would be scaled to have the same variance as these Normal errors, i.e the numerator of the scaling factor would be $\sigma_y \sqrt{(1-0.8)}$, i.e $\sqrt{(0.2)}\ \sigma_y$.

Only two values of ρ^2 were chosen for the comparison procedures, namely $\rho^2 = 0$ and $\rho^2 = 0.8$.

6.3.3 : The Error Distributions

(a) Mixture Distributions

As mentioned earlier, for Phases (a) - (c) of this set of simulations, the error distribution used was that of a mixture distribution composed of two Normal components.

For a mixture containing k Normal components, the probability density function, f(x), is given by

$$\sum_{t=1}^{k} \lambda_t \left(\sqrt{(2\Pi)}\sigma_t\right)^{-1} \exp\left[- \frac{(x-\mu_t)^2}{2\,\sigma_t^2} \right]$$

where $\lambda_1, \ldots, \lambda_t$ are known as the weights of the component distributions ($\lambda_i \geq 0$, $i = 1, \ldots, k$

$$\sum_{t=1}^{k} \lambda_t = 1 \qquad)$$

Thus for a two-component mixture, the probability density function (p.d.f.) would be

$$f(x) = \lambda\, N(\mu_1, \sigma_1^2) + (1-\lambda)\, N(\mu_2, \sigma_2^2)$$

where $N(\mu , \sigma^2)$ represents the p.d.f of a Normal distribution with mean μ and variance σ^2.

From the moment results of Johnson and Kotz(1969) it can be seen that for such a two-component mixture,

$$\mathbb{E}(X) = \lambda\mu_1 + (1-\lambda)\mu_2$$
$$\mathbb{E}(X^2) = \lambda\,(\mu_1^2 + \sigma_1^2) + (1-\lambda)\,(\mu_2^2 + \sigma_2^2)$$
$$\mathbb{E}(X^3) = \lambda\,(\mu_1^3 + 3\,\mu_1\sigma_1^2) + (1-\lambda)(\mu_2^3 + 3\,\mu_2\sigma_2^2)$$
$$\mathbb{E}(X^4) = \lambda\,(\mu_1^4 + 6\,\mu_1^2\sigma_1^2 + 3\sigma_1^4) + (1-\lambda)\,(\mu_2^4 + 6\,\mu_2^2\sigma_2^2 + 3\sigma_2^4)$$

For simplicity, it was decided to set $\lambda = {}^1/_2$ and $\mu_1 = -\mu_2 = a$, so that the simulated distributions would automatically have zero expectation, and so that the moments, above, would equal their corresponding central moments.

Also, simulations were performed using distributions with unit variance (the error variances then being scaled as necessary). Note that for a mixture distribution as defined above, and letting $\sigma_1^2 = b$ and $\sigma_2^2 = c$,

$$\mathbb{E}(X) = {}^1/_2\,(a - a) = 0$$
$$\mathbb{E}(X^2) = {}^1/_2\,(2a^2 + b + c)$$

$$\mathbb{E}(X^3) = {}^3/_2\, a(b - c)$$
$$\mathbb{E}(X^4) = {}^1/_2\, (\, 2a^4 + 6a^2(b+c) + 3(b^2 + c^2)\,)$$

so that to achieve unit variance, it would be necessary to set

$${}^1/_2\, (2a^2 + b + c) = 1 \quad ,$$

i.e $\quad 2a^2 + b + c = 2$

It was of interest to control the skewness (β) and/or the kurtosis (γ) while keeping the mean and variance fixed.

Here
$$\beta = \frac{m_3}{m_2^{3/2}} = \frac{\mathbb{E}(X^3)}{\mathbb{E}(X^2)^{3/2}}$$

and
$$\gamma = \frac{m_4}{m_2^2} = \frac{\mathbb{E}(X^4)}{\mathbb{E}(X^2)^2}$$

(b) The Lognormal Distribution

For phase (d) of the simulations, various lognormal distributions were used for the errors. A lognormal distribution is defined as follows :

For a random variable, X, if there exists a number, θ, such that $Z = \log(X - \theta)$ is Normally distributed, then X is said to have a lognormal distribution. If the mean and variance of Z are denoted by ξ and σ^2 respectively, then the p.d.f of X is as shown below.

$$p(x) = \begin{cases} [(x-\theta)\sigma\sqrt{(2\Pi)}]^{-1} \exp\{-{}^1/_2\,(\log(x-\theta)-\zeta)^2 / \sigma^2\} & \text{if } x \succ \theta \\ 0 & \text{otherwise} \end{cases}$$

If θ, a parameter affecting only the location of the distribution, takes value zero, then the 'two-parameter lognormal distribution' is obtained, where

$$\mathbb{E}(X) = \exp(\zeta + {}^1/_2\sigma^2)$$
$$V(X) = \exp(2\zeta)\,\omega(\omega - 1)$$
$$\beta = \sqrt{(\omega-1)}\;(\omega + 2)$$
$$\gamma = \omega^4 + 2\omega^3 + 3\omega^2 - 3 \qquad \text{where } \omega = \exp(\sigma^2).$$

It can be seen that the value of ζ does not affect β or γ . Since any simulated values could easily be scaled to give the required mean and variance, both θ and ζ were set to zero, so that

$$\mathbb{E}(X) = \exp(1/2\sigma^2)$$
$$V(X) = \omega(\omega - 1)$$
$$\beta = \sqrt{(\omega - 1)}\ (\omega + 2)$$
$$\gamma = \omega^4 + 2\omega^3 + 3\omega^2 - 3$$

6.3.4 : The Simulations

The error distributions used were of the form shown in Figures 6.28-6.37 . In each case, for the plots, the distributions have been scaled to have unit variance, and a Standard Normal Distribution is superimposed for comparison purposes.

The simulations performed were as detailed below.

(a) Fix $\beta = 0$, Vary γ (Using a mixture distribution)

Let Z represent the error variable. Setting $a = 0$ guarantees that $\beta = 0$ for a mixture distribution (using the earlier notation).

Then
$$\mathbb{E}(Z) = 0$$
$$V(Z) = 1/2\ (b + c)$$
$$\beta = 0$$
$$\gamma = 3/2(b^2 + c^2)$$

It was desired to fix the variance, V(Z), to take a value 1 (i.e $(b + c) = 2$) . To vary γ , different values of b and c were chosen such that $b + c = 2$, then γ was calculated using the formula above.

Simulations were carried out for $\gamma = 3.5$ (i.e. $b = 0.6$) and for $\gamma = 5.4$ (i.e. $b = 0.1$). The results were as shown in Tables 6.20-6.23 , Figures 6.38-6.41 and are summarised within Section 6.3.5 .

Figure 6.28 : Error Distn : Mixture , $\beta = 0$, $\gamma = 3.48$

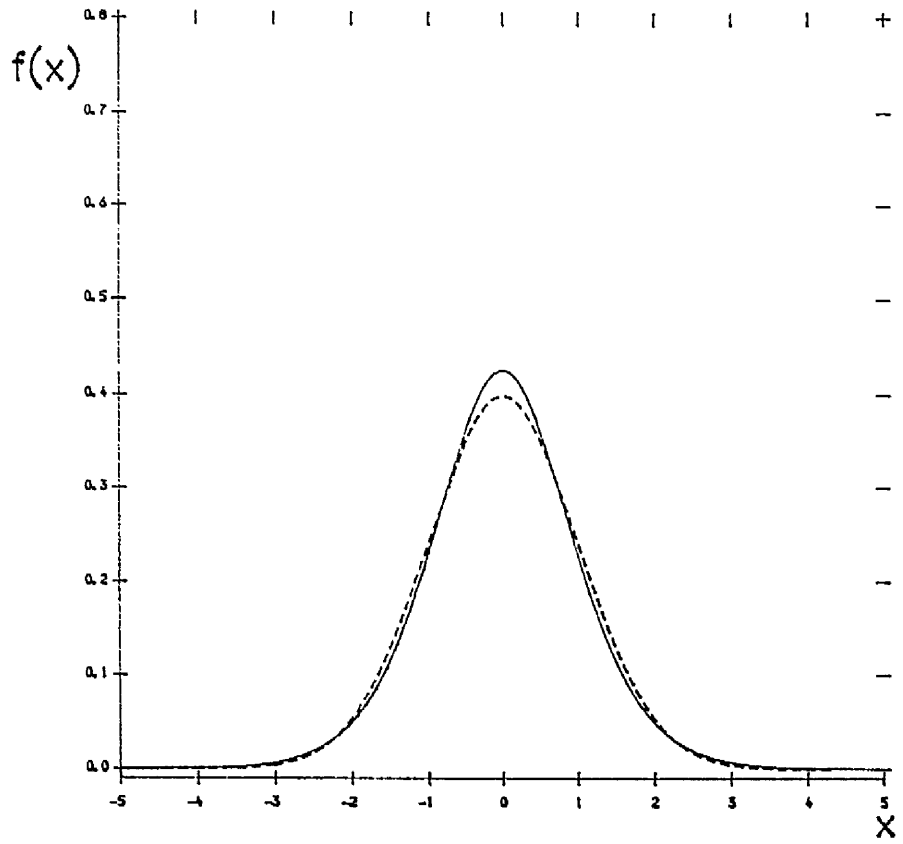


Figure 6.29 : Error Distn : Mixture , $\beta = 0$, $\gamma = 5.43$

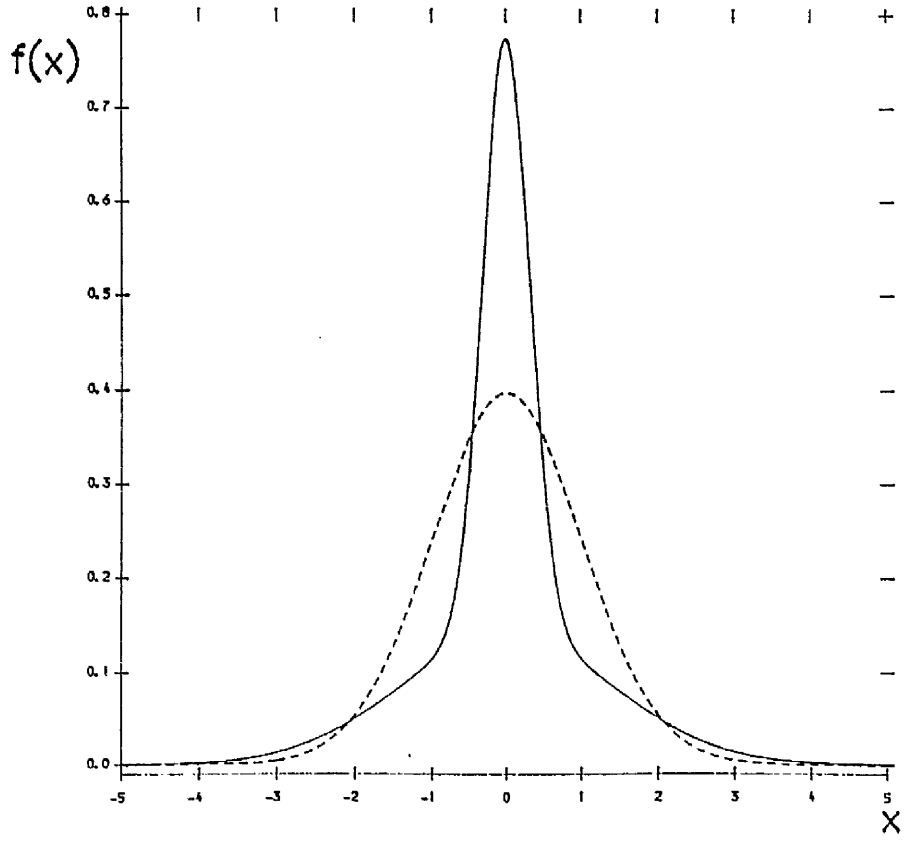


Figure 6.30 : Error Distn : Mixture , $\beta = 0.45$, $\gamma = 3$

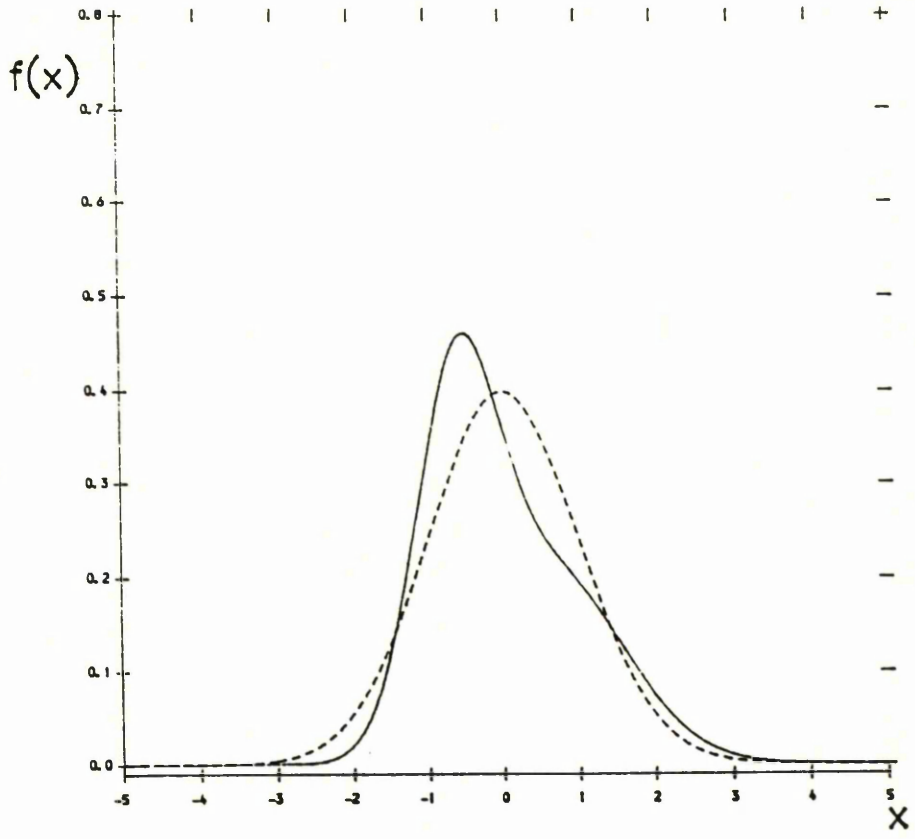


Figure 6.31 : Error Distn : Mixture , $\beta = 0.6$, $\gamma = 3$

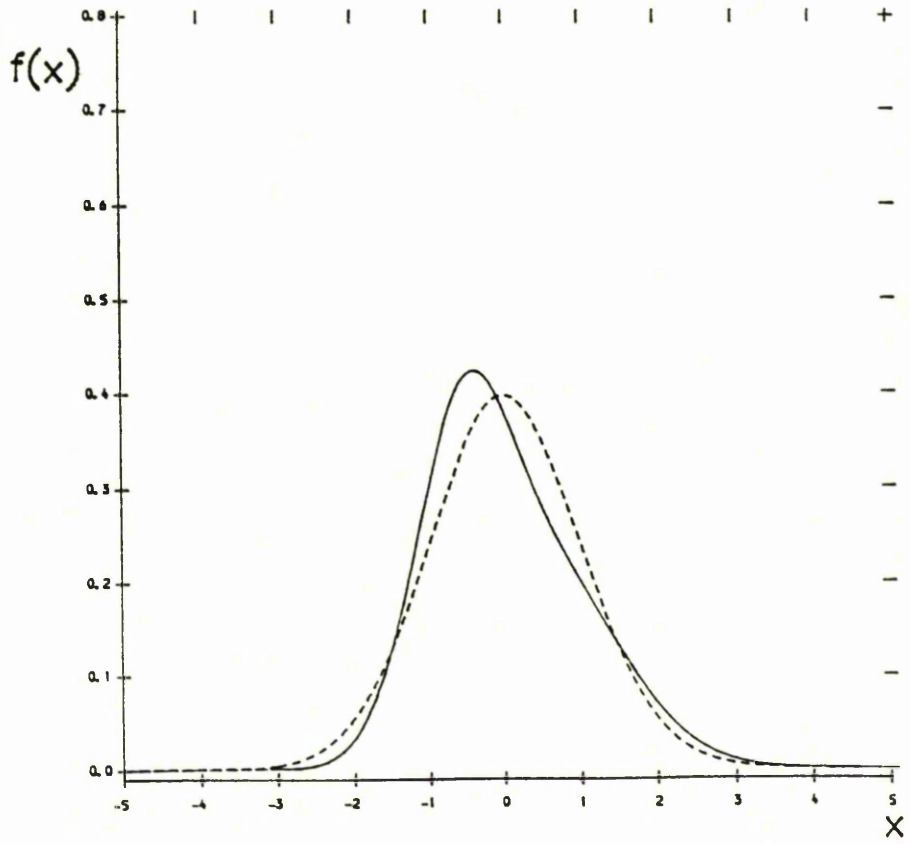


Figure 6.32 : Error Distn : Mixture , β = 0.7 , γ = 3.47

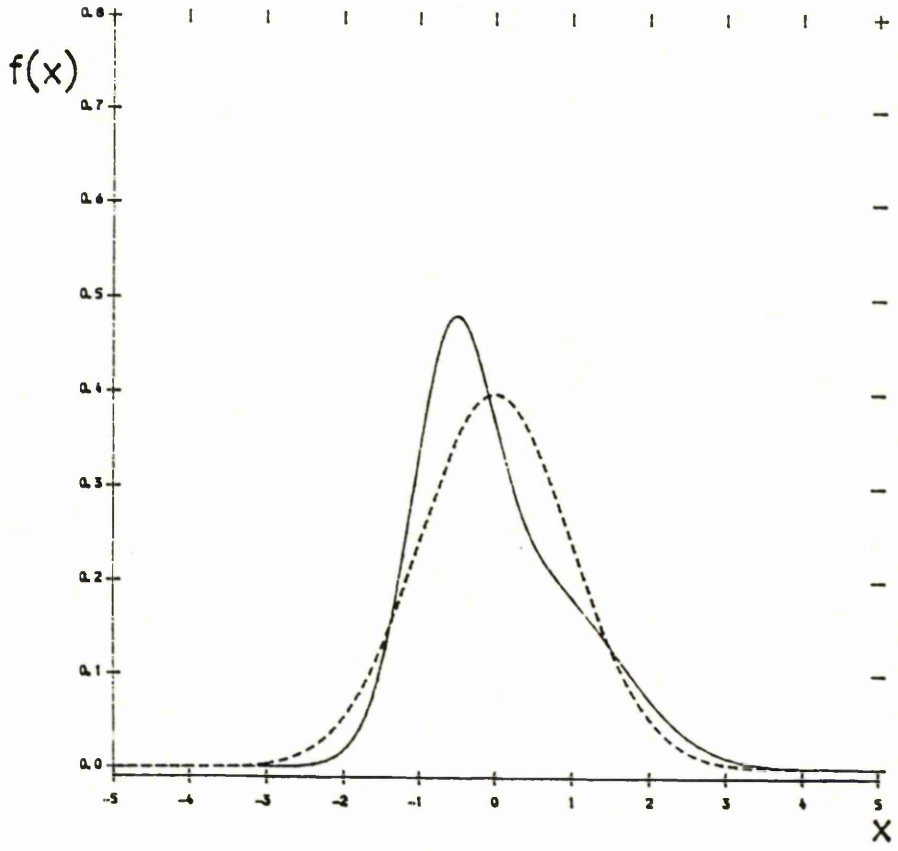


Figure 6.33 : Error Distn : Mixture , β = 4.39 , γ = 37.0

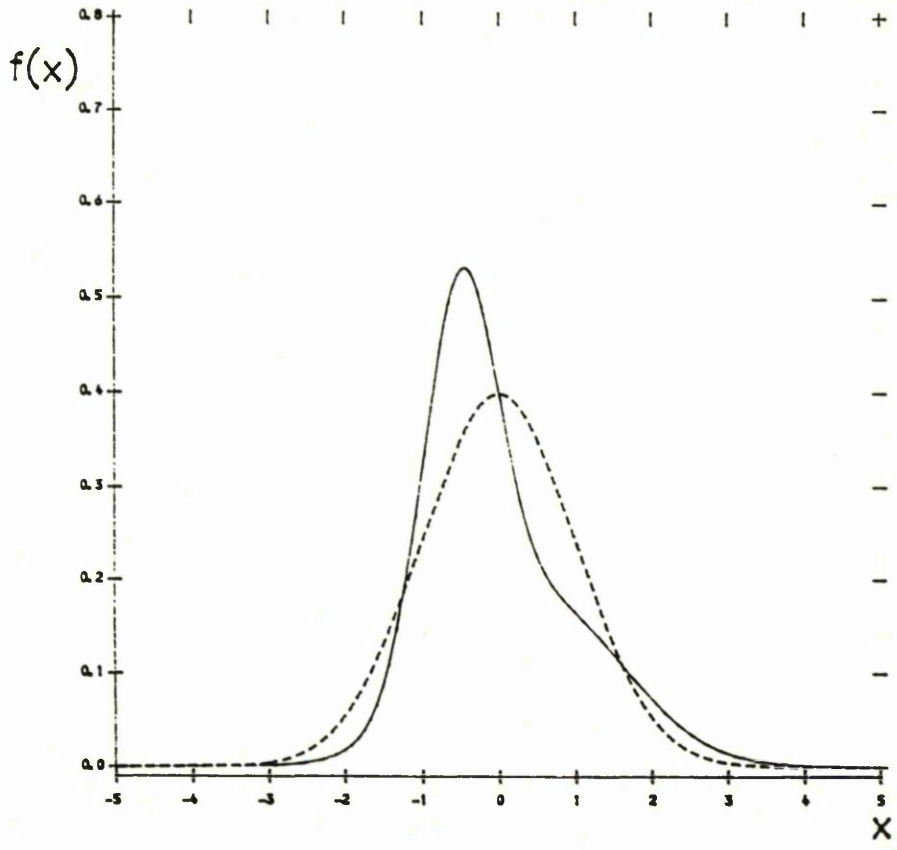


Figure 6.34 : Error Distn : Mixture , β = -4.39 , γ = 37.0

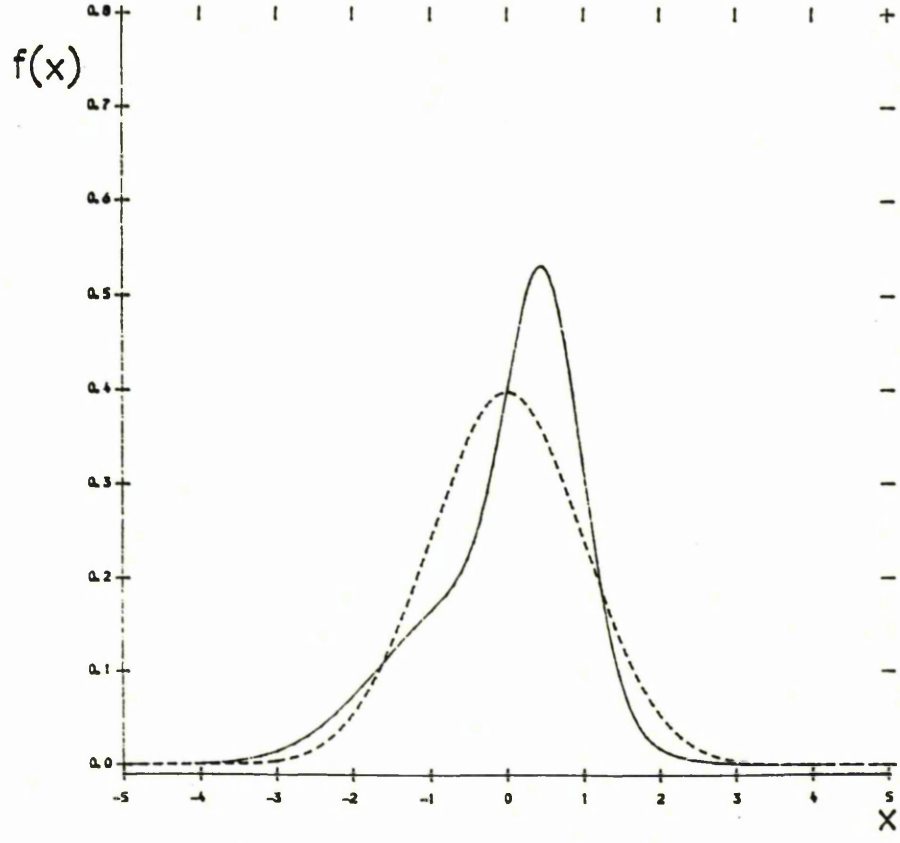


Figure 6.35 : Error Distn : Lognormal , σ = 0.3

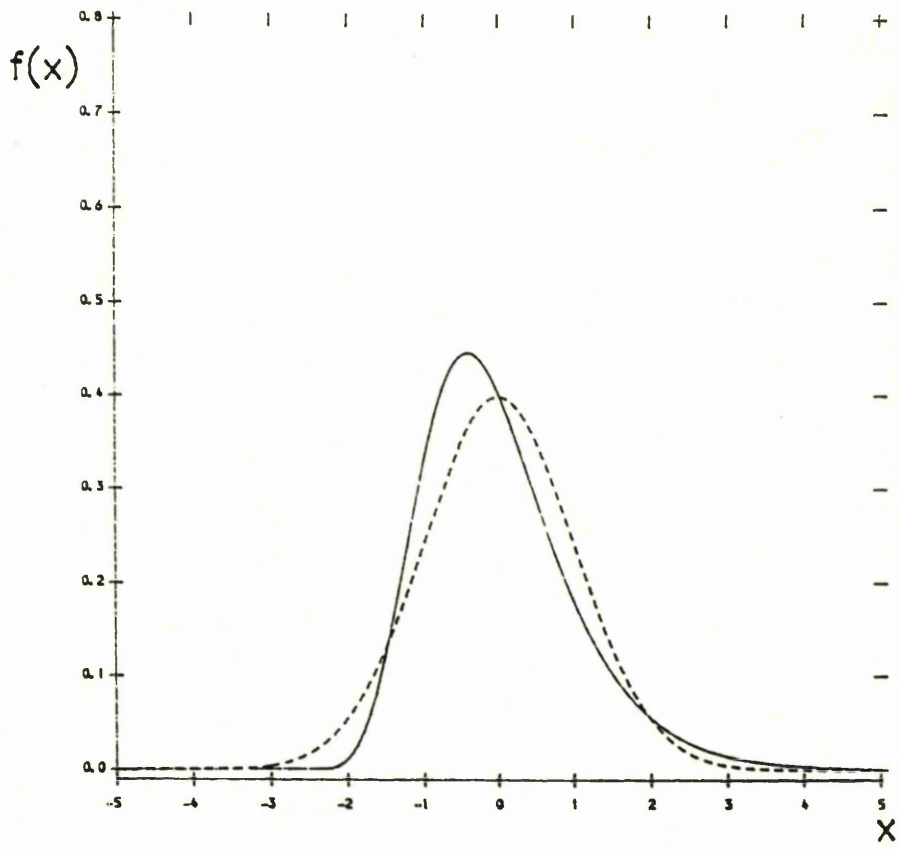


Figure 6.36 : Error Distn : Lognormal , σ = 0.4

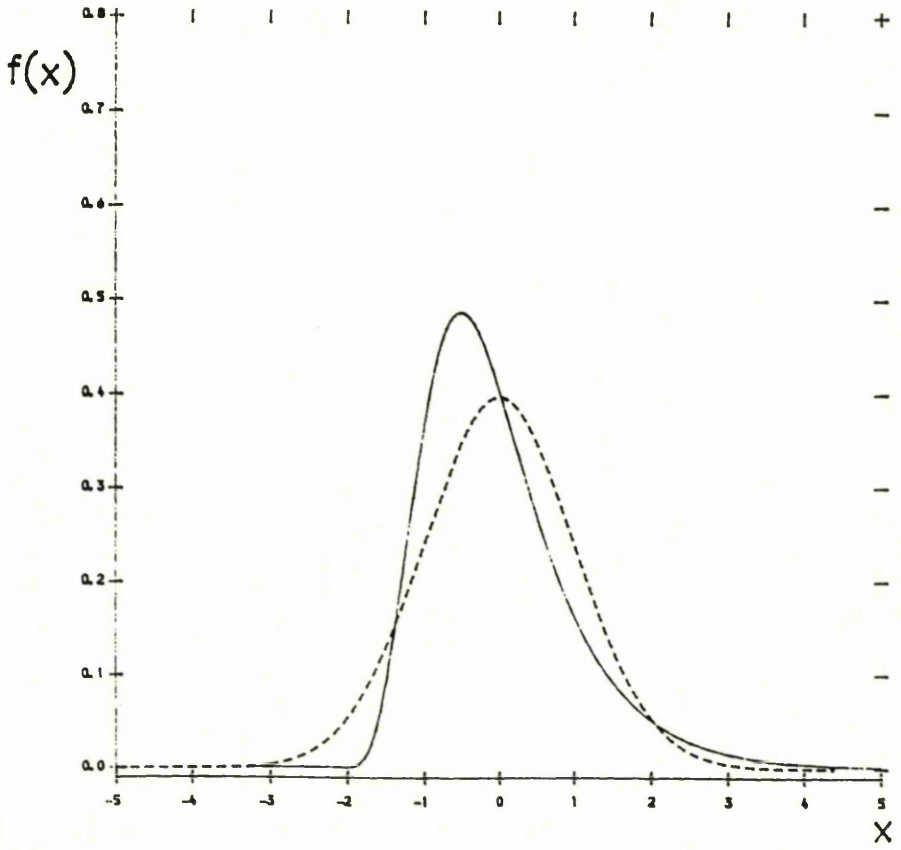


Figure 6.37 : Error Distn : Lognormal , $\sigma = 0.5$

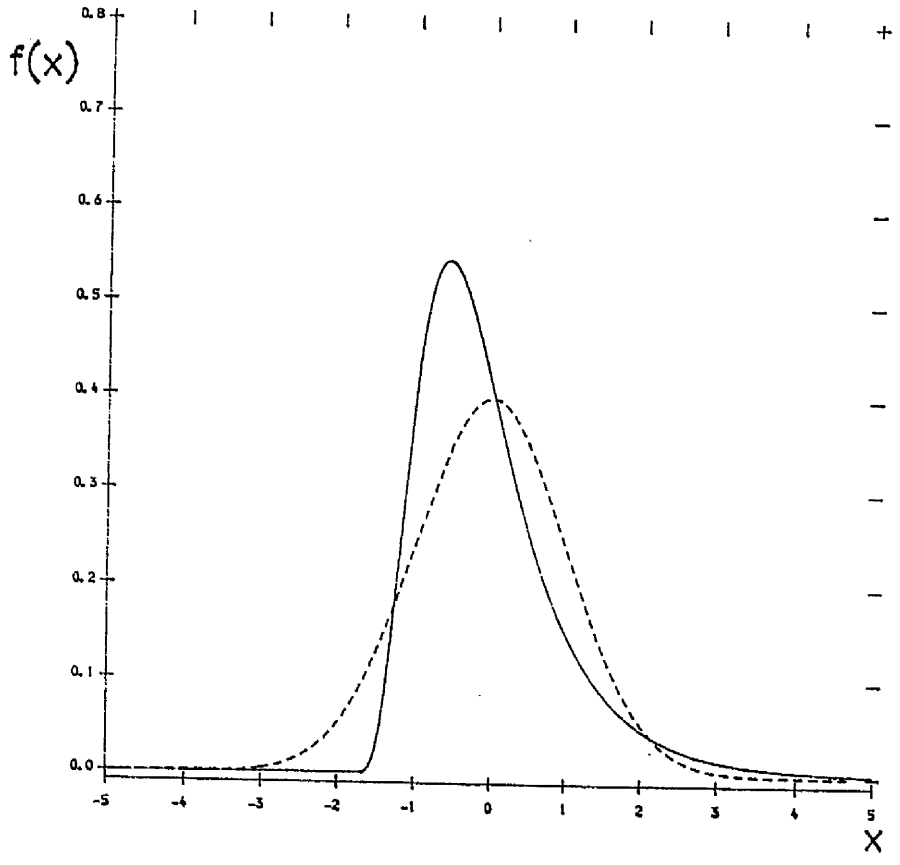


Table 6.20 : $\rho^2_{comp} = 0$, $\beta = 0$, $\gamma = 3.48$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.9	9.1	25.7	54.1	80.0	93.4	98.9	100.0	100.0
TEST 2	4.9	9.1	25.4	53.5	78.7	93.2	98.9	100.0	100.0
TEST 3	4.8	9.1	25.3	53.8	79.0	93.9	99.0	99.9	100.0
TEST 4	4.9	9.3	25.8	54.0	78.9	93.6	100.0	99.8	100.0
TEST 5	5.1	18.1	43.4	72.8	91.5	97.9	99.8	100.0	100.0
TEST 6	5.5	18.7	44.2	73.5	91.4	98.0	99.8	100.0	100.0
TEST 7	4.9	18.5	46.3	76.0	93.1	98.5	99.9	100.0	100.0
TEST 8	5.3	16.9	40.4	67.9	87.2	96.5	99.4	99.9	100.0

Table 6.21 : $\rho^2_{comp} = 0.80$, $\beta = 0$, $\gamma = 3.48$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.9	9.9	26.1	53.1	79.2	94.0	98.9	99.8	100.0
TEST 2	4.7	31.4	87.3	99.7	100.0	100.0	100.0	100.0	100.0
TEST 3	4.7	9.5	24.3	50.2	76.2	92.4	98.4	99.8	100.0
TEST 4	4.4	26.6	80.4	99.0	100.0	100.0	100.0	100.0	100.0
TEST 5	4.8	18.3	43.8	72.1	90.9	98.3	99.8	100.0	100.0
TEST 6	5.1	52.1	95.6	100.0	100.0	100.0	100.0	100.0	100.0
TEST 7	4.7	18.5	45.0	73.6	91.6	98.5	99.9	99.9	100.0
TEST 8	4.7	45.1	90.7	99.6	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Table 6.22 : $\rho^2_{comp} = 0$, $\beta = 0$, $\gamma = 5.43$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.5	10.3	26.5	55.0	79.2	93.2	98.4	99.7	100.0
TEST 2	4.5	10.2	26.0	54.6	78.7	93.0	98.2	99.7	100.0
TEST 3	4.5	15.1	44.0	76.1	92.1	97.9	99.5	99.5	100.0
TEST 4	4.7	15.3	43.9	76.1	91.7	97.8	99.5	99.9	100.0
TEST 5	4.4	19.1	44.7	73.7	90.7	97.9	99.6	100.0	100.0
TEST 6	4.8	19.7	45.7	73.9	90.7	97.9	99.6	100.0	100.0
TEST 7	4.8	29.7	68.7	91.7	98.1	99.7	100.0	100.0	100.0
TEST 8	5.1	26.0	59.1	84.9	94.6	98.7	99.8	99.9	100.0

Table 6.23 : $\rho^2_{comp} = 0.80$, $\beta = 0$, $\gamma = 5.43$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.7	10.0	27.1	52.9	78.9	93.8	99.1	99.9	100.0
TEST 2	5.0	32.9	86.2	99.6	100.0	100.0	100.0	100.0	100.0
TEST 3	4.7	9.3	25.7	51.1	76.9	92.5	98.9	99.8	100.0
TEST 4	4.7	30.2	81.5	98.9	100.0	100.0	100.0	100.0	100.0
TEST 5	4.8	18.3	44.0	72.5	91.5	98.1	99.8	100.0	100.0
TEST 6	5.5	53.5	95.1	99.9	100.0	100.0	100.0	100.0	100.0
TEST 7	4.8	19.0	45.2	74.4	92.2	98.3	99.9	100.0	100.0
TEST 8	4.9	57.8	95.5	99.8	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Figure 6.38 : $\rho^2_{comp} = 0$, $\beta = 0$, $\gamma = 3.48$

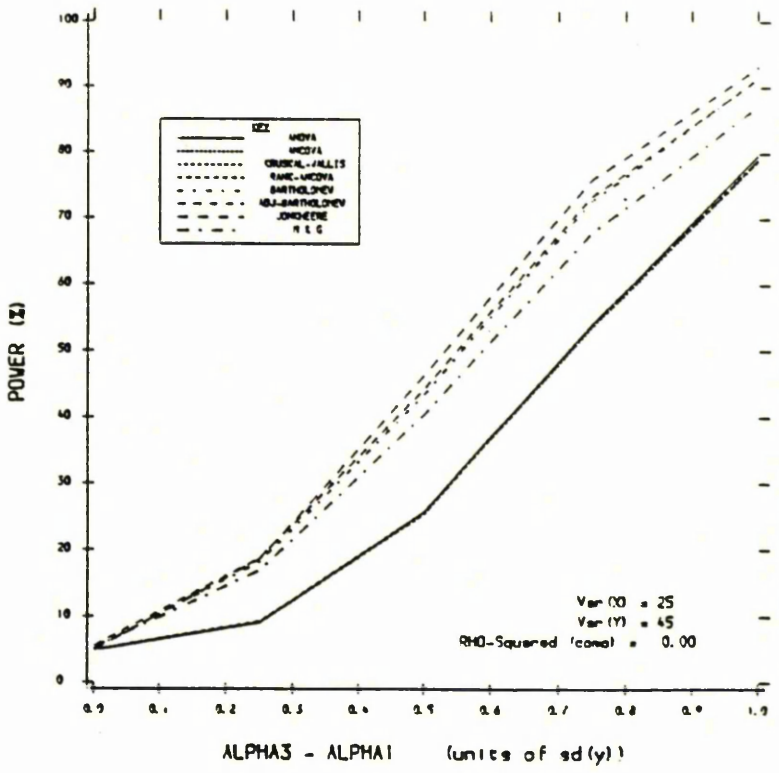


Figure 6.39 : $\rho^2_{comp} = 0.80$, $\beta = 0$, $\gamma = 3.48$

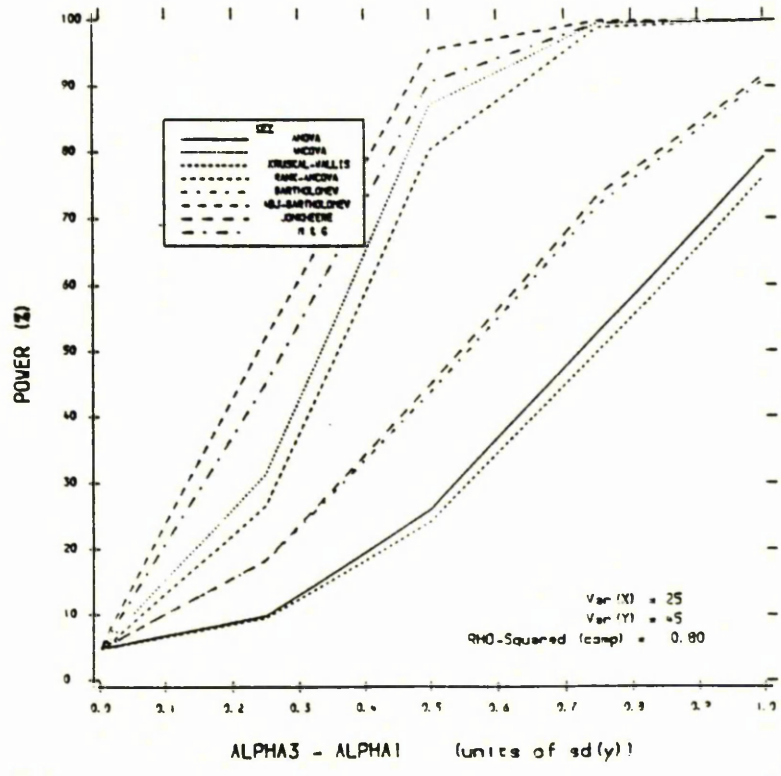


Figure 6.40 : $\rho^2_{comp} = 0$, $\beta = 0$, $\gamma = 5.43$

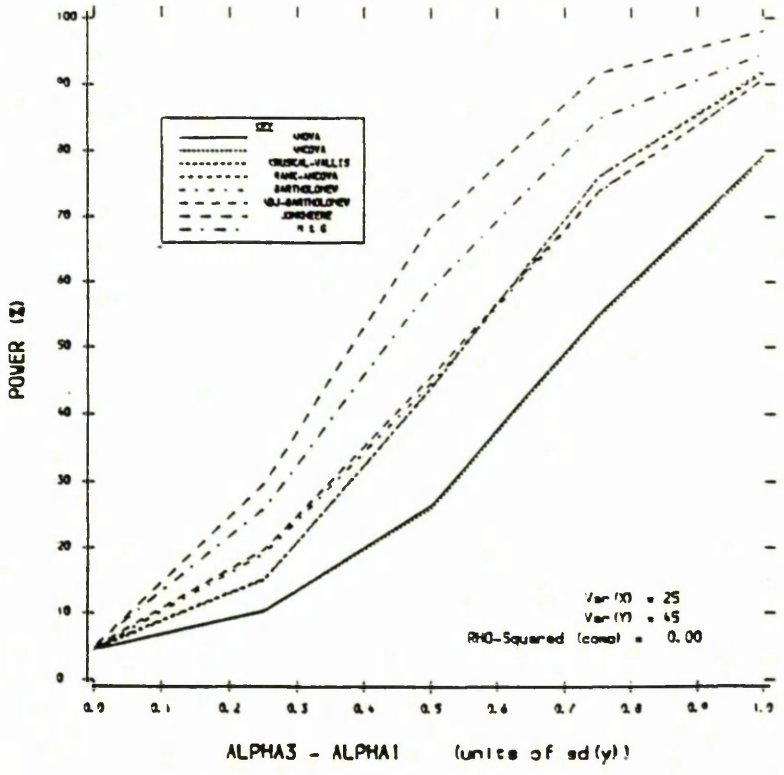
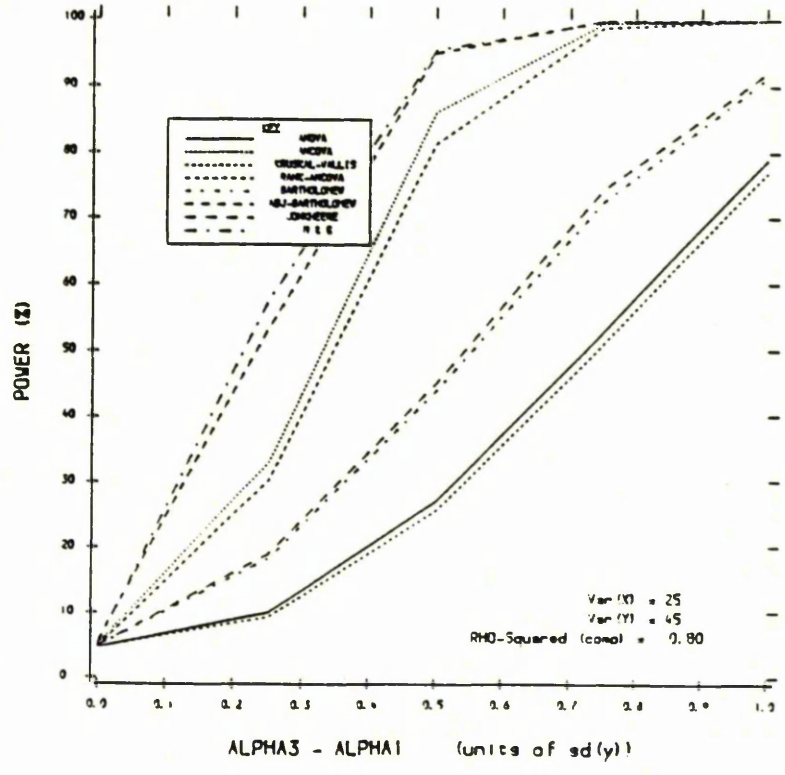


Figure 6.41 : $\rho^2_{comp} = 0.80$, $\beta = 0$, $\gamma = 5.43$



(b) Fix $\gamma = 3$, Vary β (Using mixture distributions)

For unit variance, $2a^2 + b + c = 2$ (using the earlier notation).

The expression for γ was

$$\gamma = a^4 + 3a^2(b+c) + {}^3/_2(b^2 + c^2) = 3 \text{ (in this case).} \quad (6.1)$$

Substituting for $a^2 = {}^1/_2(2-b-c)$ in equation (6.1) gives

$${}^1/_4(2 - b - c)^2 + {}^3/_2(2 - b - c)(b + c) + {}^3/_2(b^2 + c^2) = 3$$

i.e $b^2 + c^2 + 8b + 8c - 10bc = 8$ (after some algebra)

Completing the square in b,

$$(b - 5c + 4)^2 = 24\,(c - 1)^2$$

i.e $$b = 5c - 4 \pm 2\sqrt{6}\,(c-1)$$

To vary β while constraining γ to take value 3, the procedure followed was as follows:

(i) Choose $\sigma_2^2 = c$

(ii) Calculate $b = 5c - 4 \pm 2\sqrt{6}\,(c-1)$ and choose the value for b (or a value for b) giving both b and (2-b-c) as non-negative.

(iii) Calculate $a = \sqrt{[\ {}^1/_2\ (2-b-c)\ ]}$

(iv) Calculate β

Simulations were performed for $\beta = 0.45$ ($c = 0.41$) and $\beta = 0.59$ ($c = 0.30$). The results were as shown in Tables 6.24-6.27, Figures 6.42-6.45 , and summarised within Section 6.3.5.

One further mixture distribution was chosen for this section, to give a skewness more severe than could have been achieved while constraining $\gamma = 3$, but the kurtosis not permitted to exceed the lower value used in simulations (a). The chosen distribution was one with $\beta = 0.72$, $\gamma = 3.5$. It was intended to compare these results to those obtained from the rest of Phase (b), essentially disregarding the small discrepancy between the kurtosis used and that of a Normal distribution. The results are

Table 6.24 : $\rho^2_{comp} = 0$, $\beta = 0.45$, $\gamma = 3$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.1	8.8	26.2	53.2	80.0	94.3	98.9	99.9	100.0
TEST 2	5.2	8.8	25.8	52.4	79.1	94.1	98.9	99.9	100.0
TEST 3	5.1	8.5	25.5	51.3	78.5	93.8	98.5	99.9	100.0
TEST 4	5.5	8.8	26.2	51.5	78.4	93.5	98.3	99.9	100.0
TEST 5	5.1	17.8	43.9	72.2	91.6	98.1	99.8	100.0	100.0
TEST 6	5.4	18.7	44.3	72.5	91.7	98.2	99.7	100.0	100.0
TEST 7	5.0	18.9	45.9	74.5	92.7	98.6	99.8	100.0	100.0
TEST 8	5.4	17.3	40.1	66.6	86.7	96.1	99.0	99.9	100.0

Table 6.25 : $\rho^2_{comp} = 0.80$, $\beta = 0.45$, $\gamma = 3$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.8	9.8	26.1	53.3	79.0	94.3	99.0	99.9	100.0
TEST 2	5.1	31.7	87.7	99.8	100.0	100.0	100.0	100.0	100.0
TEST 3	5.3	9.2	24.0	50.2	76.0	93.0	98.5	99.8	100.0
TEST 4	5.2	27.2	80.2	98.9	100.0	100.0	100.0	100.0	100.0
TEST 5	5.5	18.5	43.9	72.0	91.2	98.3	99.9	100.0	100.0
TEST 6	5.5	51.8	95.8	100.0	100.0	100.0	100.0	100.0	100.0
TEST 7	5.1	18.6	45.2	74.2	92.0	98.3	99.9	100.0	100.0
TEST 8	4.9	44.4	91.0	99.6	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Table 6.26 : $\rho^2_{comp} = 0$, $\beta = 0.6$, $\gamma = 3$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.1	9.5	26.2	53.0	79.1	94.1	98.9	99.9	100.0
TEST 2	5.0	9.6	25.6	52.5	78.1	93.6	98.7	99.9	100.0
TEST 3	5.1	9.6	27.2	53.5	79.2	93.7	98.5	99.9	100.0
TEST 4	5.2	9.8	27.4	53.8	78.9	93.4	98.5	99.8	100.0
TEST 5	5.0	18.1	43.8	72.2	91.6	98.3	99.8	100.0	100.0
TEST 6	5.4	19.1	44.4	72.8	91.7	98.4	99.8	100.0	100.0
TEST 7	5.0	20.0	47.8	76.7	94.0	98.8	99.8	100.0	100.0
TEST 8	5.5	18.0	41.1	68.1	87.7	96.5	99.2	99.9	100.0

Table 6.27 : $\rho^2_{comp} = 0.80$, $\beta = 0.6$, $\gamma = 3$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.8	9.9	25.4	52.8	78.9	94.9	99.0	99.9	100.0
TEST 2	4.7	30.7	86.7	99.8	100.0	100.0	100.0	100.0	100.0
TEST 3	4.5	9.4	24.2	50.3	76.5	92.9	98.8	100.0	100.0
TEST 4	5.0	26.2	79.2	100.0	100.0	100.0	100.0	100.0	100.0
TEST 5	5.1	18.8	43.2	71.7	91.0	98.4	100.0	100.0	100.0
TEST 6	5.2	51.1	95.4	100.0	100.0	100.0	100.0	100.0	100.0
TEST 7	5.0	19.0	44.5	72.9	91.7	98.6	99.9	100.0	100.0
TEST 8	4.8	44.9	90.7	99.7	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Figure 6.42 : $\rho^2_{comp} = 0$, $\beta = 0.45$, $\gamma = 3$

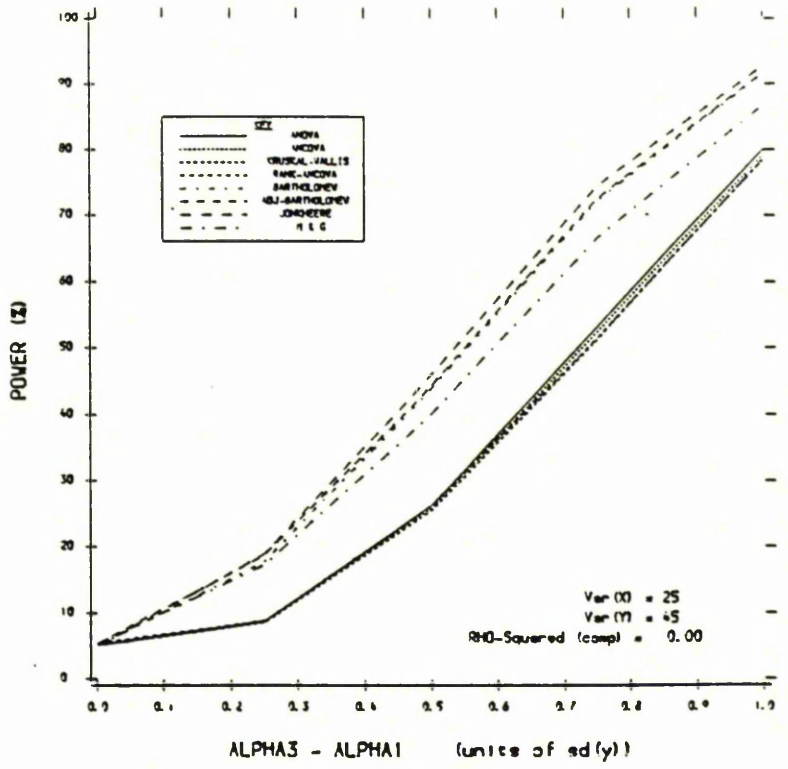


Figure 6.43 : $\rho^2_{comp} = 0.80$, $\beta = 0.45$, $\gamma = 3$

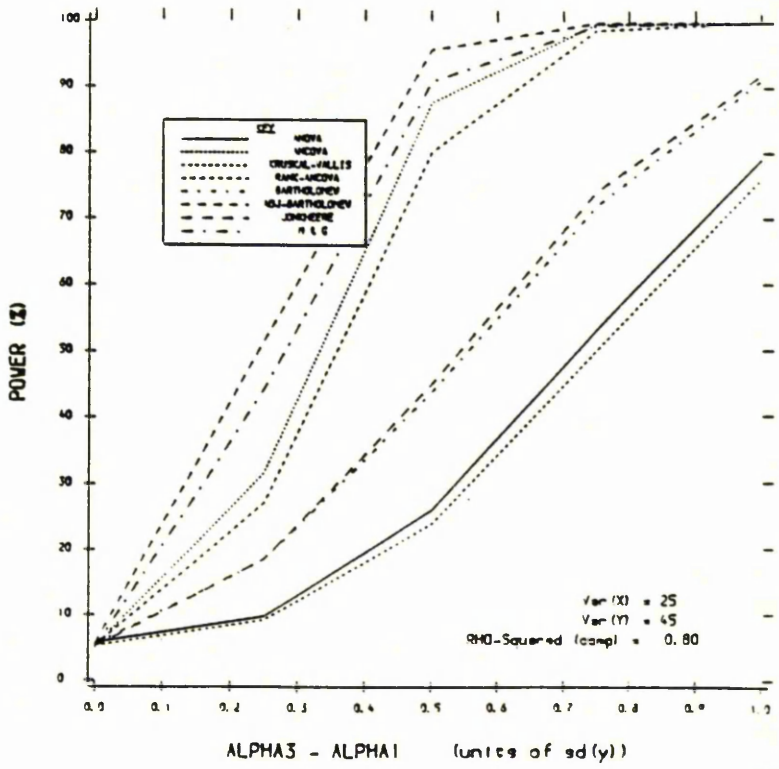


Figure 6.44 : $\rho^2_{comp} = 0$, $\beta = 0.6$, $\gamma = 3$

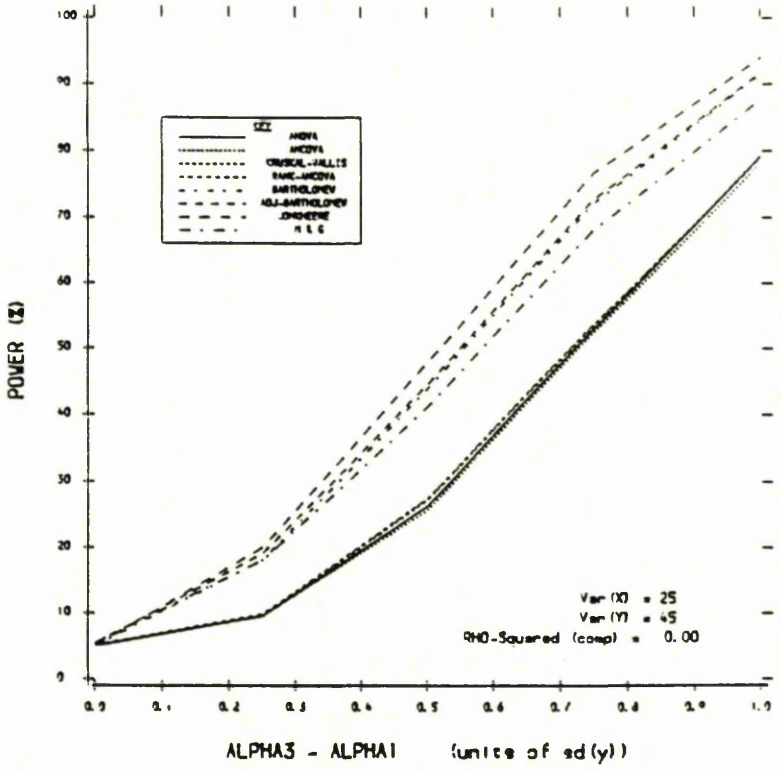
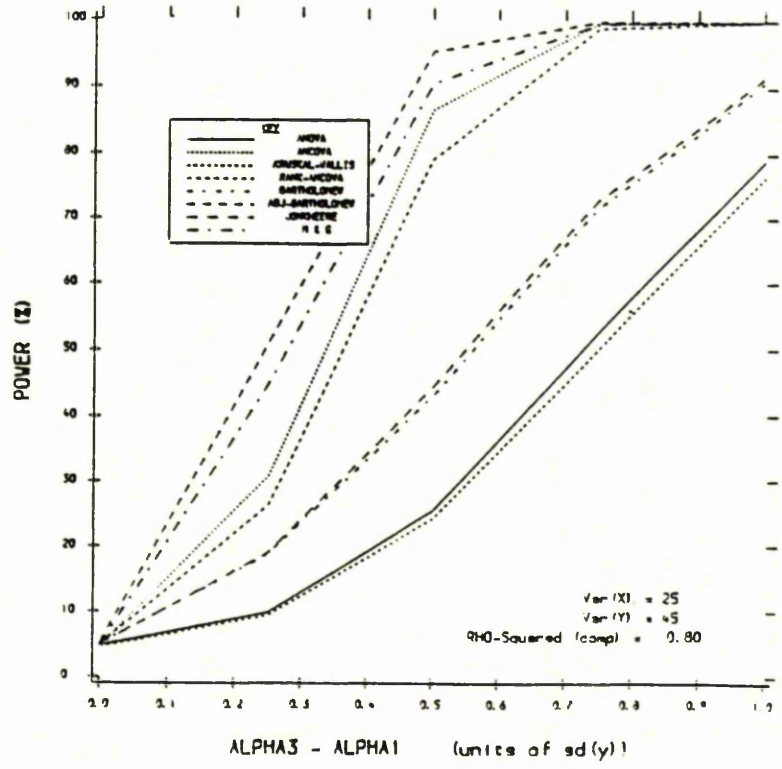


Figure 6.45 : $\rho^2_{comp} = 0.80$, $\beta = 0.6$, $\gamma = 3$



shown in Tables 6.28 & 6.29 , Figures 6.46 & 6.47 and summarised within Section 6.3.5 .

(c) Assessing the Effect of Reversing the Skewness of the Error Distribution .

Here the results of using two severely non-Normal, oppositely skewed, mixture distributions were compared (one of these being the exact mirror-image of the other, reflecting over the line $X = 0$), to confirm that the results were invariant with respect to sign of skewness.

The distributions both had zero mean, the same variance, $\gamma = 37.0$, while one had a skewness of 4.4 and the other had a skewness of -4.4 .

These results are shown in Tables 6.30 & 6.31 , Figures 6.48 & 6.49 , and are summarised within Section 6.3.5 . Note that here, a larger number of simulations was carried out than previously, in order to make any differences between the results more evident.

(d) Simultaneously varying β and γ

For the Lognormal distribution, as defined earlier, the skewness and kurtosis were varied by altering the parameter σ. The chosen values were $\sigma = 0.3, 0.4, 0.5$, leading to (β , γ) combinations of $(0.95 , 4.6)$, $(1.32 , 6.2)$ and $(1.75 , 8.9)$ respectively. These results are shown in Tables 6.32-6.37 , Figures 6.50-6.55 , and are summarised within Section 6.3.5 , below.

6.3.5 : The Results

(1) In general, incorporating covariate information (in the case of non-zero 'correlation'), and/or ordering information about the group means, increased the power of the tests. Also, in the presence of Normal errors, Normal tests tended to perform better than their non-parametric counterparts. Two noticeable exceptions were :

(i) Marcus and Genizi's test (Test 8) performed unexpectedly poorly, compared to what would be expected from the pattern of the other results. In fact, almost every 'unexpected' result for

Table 6.28 : $\rho^2_{comp} = 0$, $\beta = 0.7$, $\gamma = 3.6$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.8	9.4	26.3	53.0	78.9	94.3	99.2	100.0	100.0
TEST 2	4.8	9.2	26.1	52.4	78.0	93.7	99.2	99.9	100.0
TEST 3	4.7	9.6	27.9	55.6	79.2	93.6	98.9	99.9	100.0
TEST 4	5.1	10.0	28.6	55.9	79.4	93.6	98.9	99.9	100.0
TEST 5	4.7	18.1	43.8	72.7	91.2	98.4	99.9	100.0	100.0
TEST 6	5.1	18.7	44.7	73.3	91.4	98.4	99.9	100.0	100.0
TEST 7	4.8	20.5	49.5	79.1	93.6	98.9	99.9	100.0	100.0
TEST 8	5.4	18.7	44.0	69.9	87.7	96.5	99.2	99.9	100.0

Table 6.29 : $\rho^2_{comp} = 0.80$, $\beta = 0.7$, $\gamma = 3.6$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.3	9.9	25.4	53.8	79.7	94.3	98.9	99.9	100.0
TEST 2	5.2	32.0	87.4	99.8	100.0	100.0	100.0	100.0	100.0
TEST 3	4.9	9.6	23.9	50.9	77.1	92.8	98.4	100.0	100.0
TEST 4	4.9	26.9	80.2	99.0	100.0	100.0	100.0	100.0	100.0
TEST 5	5.3	18.7	43.6	73.0	91.9	98.2	99.8	100.0	100.0
TEST 6	5.5	52.0	95.5	99.9	100.0	100.0	100.0	100.0	100.0
TEST 7	5.2	19.3	44.7	74.1	92.2	98.4	99.7	100.0	100.0
TEST 8	5.0	46.4	91.0	99.6	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Figure 6.46 : $\rho^2_{comp} = 0$, $\beta = 0.7$, $\gamma = 3.6$

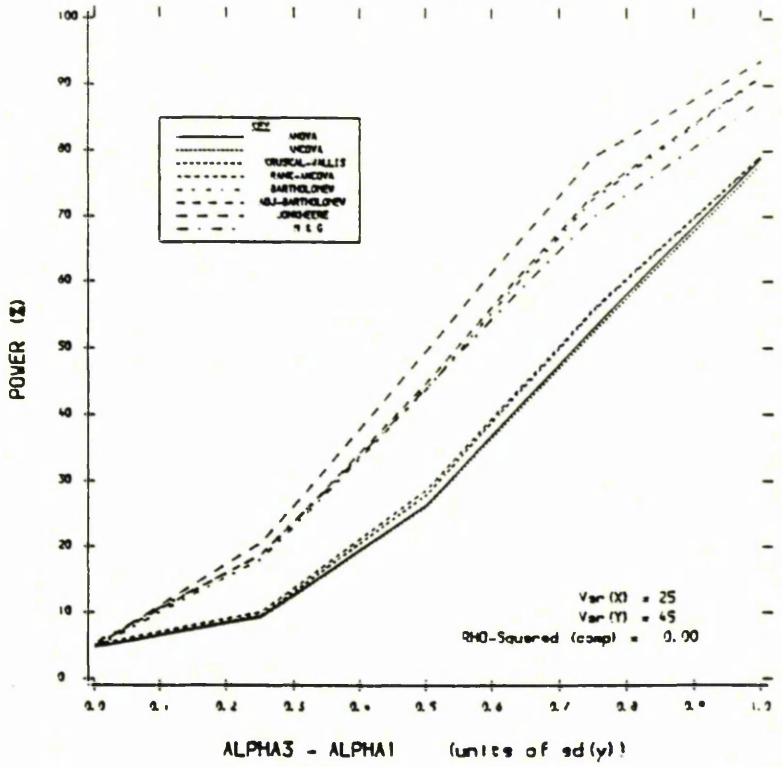


Figure 6.47 : $\rho^2_{comp} = 0.80$, $\beta = 0.7$, $\gamma = 3.6$

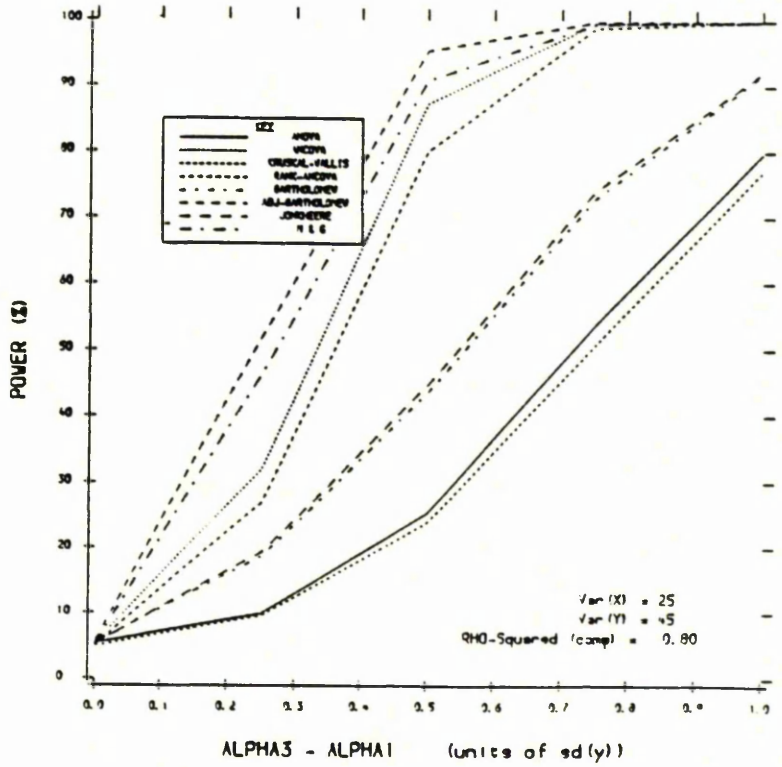


Table 6.30 : $\rho^2_{comp} = 0$, $\beta = 4.4$, $\gamma = 37.0$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.0	9.9	26.3	53.8	79.2	94.2	98.9	99.9	100.0
TEST 2	5.0	9.6	26.1	53.0	78.5	93.8	98.7	99.9	100.0
TEST 3	4.9	10.7	30.6	60.0	84.0	95.7	99.2	99.9	100.0
TEST 4	5.2	11.0	31.0	60.4	83.9	95.5	99.1	99.9	100.0
TEST 5	4.7	18.7	43.9	72.6	91.4	98.2	99.7	100.0	100.0
TEST 6	5.1	19.1	44.9	73.4	91.7	98.2	99.8	100.0	100.0
TEST 7	4.7	21.8	53.2	81.8	95.2	99.2	99.9	100.0	100.0
TEST 8	5.1	19.7	46.8	73.7	90.4	97.4	99.5	99.9	100.0

Table 6.31 : $\rho^2_{comp} = 0$, $\beta = -4.4$, $\gamma = 37.0$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.7	9.6	26.8	53.5	79.4	93.9	98.4	99.9	100.0
TEST 2	4.8	9.6	26.5	52.4	78.6	93.5	98.8	99.9	100.0
TEST 3	4.6	10.5	31.5	60.1	83.9	95.6	99.2	99.9	100.0
TEST 4	5.0	10.9	31.8	60.2	83.6	95.5	99.1	99.9	100.0
TEST 5	4.5	18.4	44.5	72.4	91.0	98.0	99.8	100.0	100.0
TEST 6	4.9	19.3	45.1	73.1	91.1	98.1	99.8	100.0	100.0
TEST 7	4.8	21.7	53.8	81.5	95.5	99.3	99.9	100.0	100.0
TEST 8	5.1	19.8	46.9	73.1	90.2	97.2	99.5	99.9	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Figure 6.48 : $\rho^2_{comp} = 0$, $\beta = 4.4$, $\gamma = 37.0$

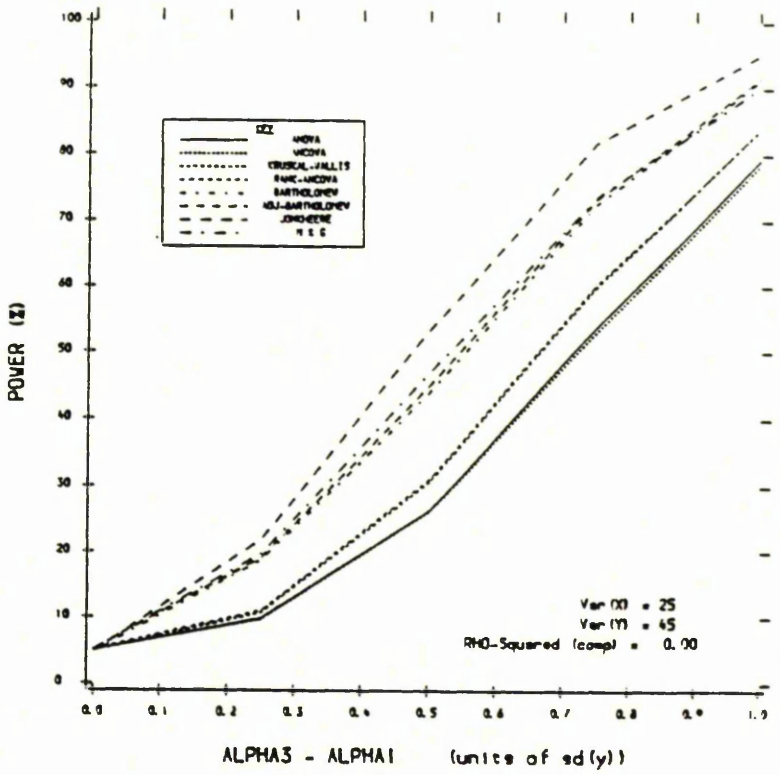


Figure 6.49 : $\rho^2_{comp} = 0$, $\beta = -4.4$, $\gamma = 37.0$

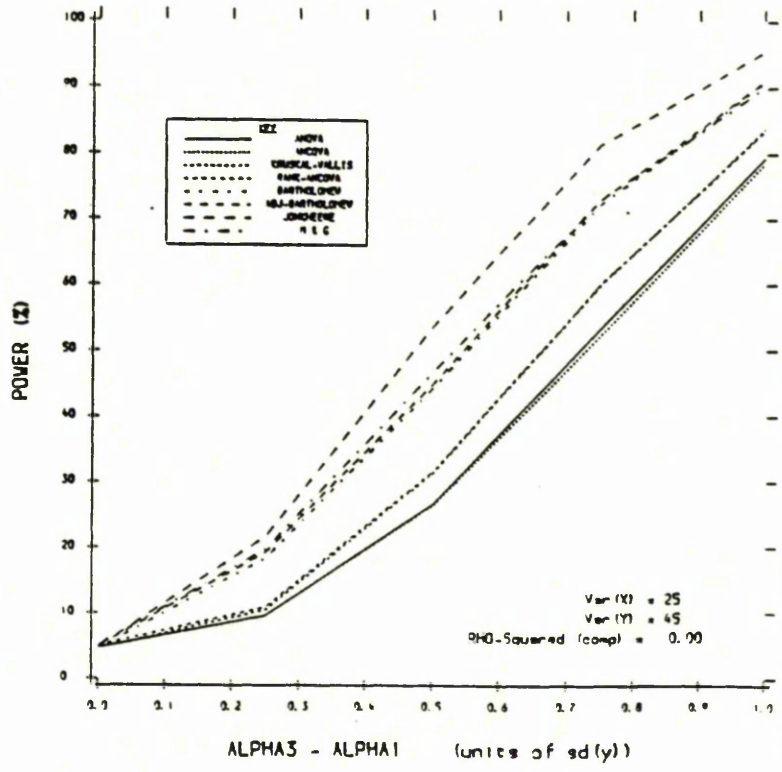


Table 6.32 : $\rho^2_{comp} = 0$. Lognormal . $\sigma = 0.3$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.7	9.7	26.6	54.3	80.3	93.5	98.8	99.8	100.0
TEST 2	4.6	9.5	26.0	53.9	79.3	93.0	98.7	99.7	100.0
TEST 3	4.4	9.9	28.7	57.8	83.3	95.4	99.4	99.9	100.0
TEST 4	4.9	10.0	29.3	57.8	83.3	95.1	99.4	99.9	100.0
TEST 5	4.8	18.3	44.0	72.4	91.0	97.8	99.8	99.9	100.0
TEST 6	5.0	19.0	45.2	72.9	91.2	97.7	99.8	99.9	100.0
TEST 7	4.9	20.3	50.2	79.8	95.1	99.1	99.9	100.0	100.0
TEST 8	5.1	19.2	43.9	71.5	89.9	96.9	99.7	99.9	100.0

Table 6.33 : $\rho^2_{comp} = 0.80$. Lognormal . $\sigma = 0.3$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	5.3	9.5	25.9	53.6	79.6	94.1	98.8	99.9	100.0
TEST 2	4.8	31.7	87.1	99.6	100.0	100.0	100.0	100.0	100.0
TEST 3	5.0	9.2	24.1	50.8	77.1	92.6	98.3	99.8	100.0
TEST 4	4.7	27.6	81.1	98.9	100.0	100.0	100.0	100.0	100.0
TEST 5	5.3	18.4	43.3	72.1	92.0	98.2	99.7	100.0	100.0
TEST 6	5.6	52.5	95.7	99.9	100.0	100.0	100.0	100.0	100.0
TEST 7	5.3	18.7	45.0	73.8	92.7	98.4	99.8	100.0	100.0
TEST 8	5.2	47.5	92.8	99.8	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Table 6.34 : $\rho^2_{comp} = 0$, Lognormal , $\sigma = 0.4$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.5	9.8	28.0	54.4	79.1	93.5	98.5	99.8	99.9
TEST 2	4.4	9.5	27.6	53.5	78.2	93.2	98.4	99.7	99.9
TEST 3	4.3	10.7	32.7	63.3	87.5	97.3	99.6	100.0	100.0
TEST 4	4.5	11.0	33.1	63.5	87.1	97.1	99.5	100.0	100.0
TEST 5	5.3	18.8	45.5	72.9	91.3	97.8	99.7	100.0	100.0
TEST 6	5.6	19.4	46.4	73.5	91.1	98.0	99.7	100.0	100.0
TEST 7	5.2	22.8	55.4	84.1	96.9	99.7	100.0	100.0	100.0
TEST 8	5.5	20.4	49.2	76.3	92.1	98.3	99.8	100.0	100.0

Table 6.35 : $\rho^2_{comp} = 0.80$, Lognormal , $\sigma = 0.4$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.5	9.5	26.5	53.7	79.7	93.5	99.0	99.9	100.0
TEST 2	5.2	32.9	87.0	99.3	100.0	100.0	100.0	100.0	100.0
TEST 3	4.5	9.1	24.8	51.4	77.4	92.6	98.9	99.8	100.0
TEST 4	4.9	29.6	82.6	99.0	100.0	100.0	100.0	100.0	100.0
TEST 5	4.6	17.8	43.9	73.5	91.3	98.2	99.8	100.0	100.0
TEST 6	5.1	52.7	95.1	99.9	100.0	100.0	100.0	100.0	100.0
TEST 7	4.8	18.5	46.2	74.7	92.2	98.5	99.8	100.0	100.0
TEST 8	5.1	51.1	94.1	99.8	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Table 6.36 : $\rho^2_{comp} = 0$. Lognormal . $\sigma = 0.5$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.8	10.0	28.1	55.6	79.8	93.0	98.2	99.5	99.9
TEST 2	4.8	9.9	27.8	55.0	79.1	92.8	98.2	99.5	99.9
TEST 3	4.9	12.4	37.7	70.4	90.8	98.3	99.9	100.0	100.0
TEST 4	5.2	12.8	38.4	70.4	90.8	98.3	99.9	100.0	100.0
TEST 5	4.8	19.0	46.2	73.8	90.2	97.4	99.5	99.8	99.9
TEST 6	5.2	19.7	46.7	74.1	90.4	97.4	99.5	99.8	99.9
TEST 7	5.1	24.5	61.4	88.6	97.8	99.8	100.0	100.0	100.0
TEST 8	5.0	21.7	53.6	81.2	94.9	98.7	99.8	100.0	100.0

Table 6.37 : $\rho^2_{comp} = 0.80$. Lognormal . $\sigma = 0.5$

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
TEST 1	4.5	9.0	26.0	54.2	79.7	93.7	99.0	99.8	100.0
TEST 2	4.6	33.4	86.8	99.2	100.0	100.0	100.0	100.0	100.0
TEST 3	4.6	8.6	24.5	51.8	77.9	92.6	98.7	99.8	100.0
TEST 4	4.8	29.6	83.8	99.0	100.0	100.0	100.0	100.0	100.0
TEST 5	5.1	17.7	43.8	72.7	91.5	98.1	99.9	100.0	100.0
TEST 6	5.2	54.1	94.7	99.8	100.0	100.0	100.0	100.0	100.0
TEST 7	5.4	18.2	45.9	74.6	92.3	98.3	99.9	100.0	100.0
TEST 8	5.2	54.5	96.0	99.9	100.0	100.0	100.0	100.0	100.0

The Tests

TEST 1 : ANOVA
TEST 2 : ANCOVA
TEST 3 : KRUSKAL - WALLIS
TEST 4 : RANK ANCOVA
TEST 5 : BARTHOLOMEW
TEST 6 : BARTHOLOMEW - ADJ
TEST 7 : JONKHEERE
TEST 8 : MARCUS & GENIZI

Group Spacings (Alpha 3 - Alpha 1) (s.d. units)

(a) 0.0 (b) 0.25 (c) 0.50
(d) 0.75 (e) 1.00 (f) 1.25
(g) 1.50 (h) 1.75 (i) 2.00

Figure 6.50 : $\rho^2_{comp} = 0$, Lognormal , $\sigma = 0.3$

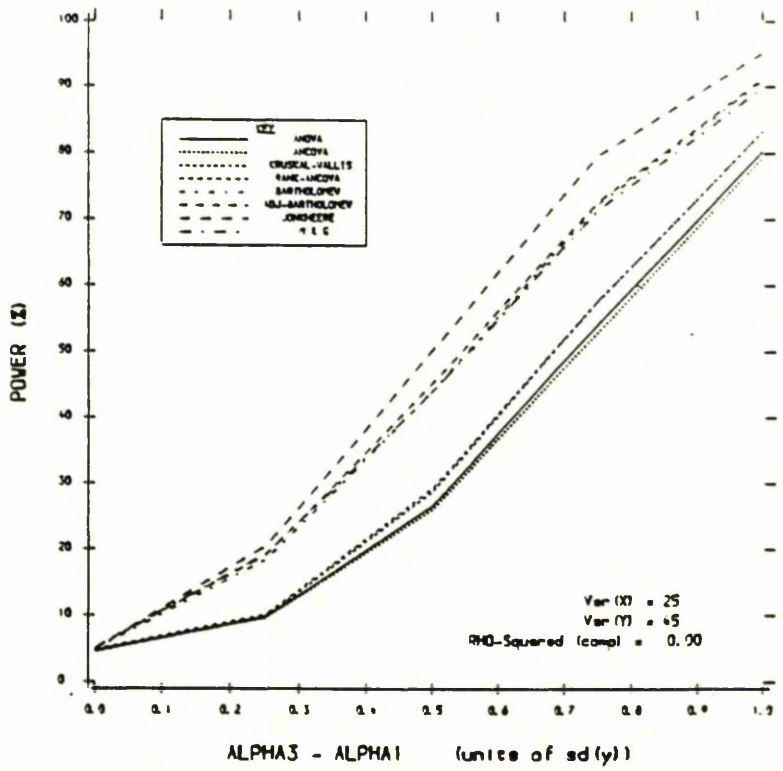


Figure 6.51 : $\rho^2_{comp} = 0.80$, Lognormal , $\sigma = 0.3$

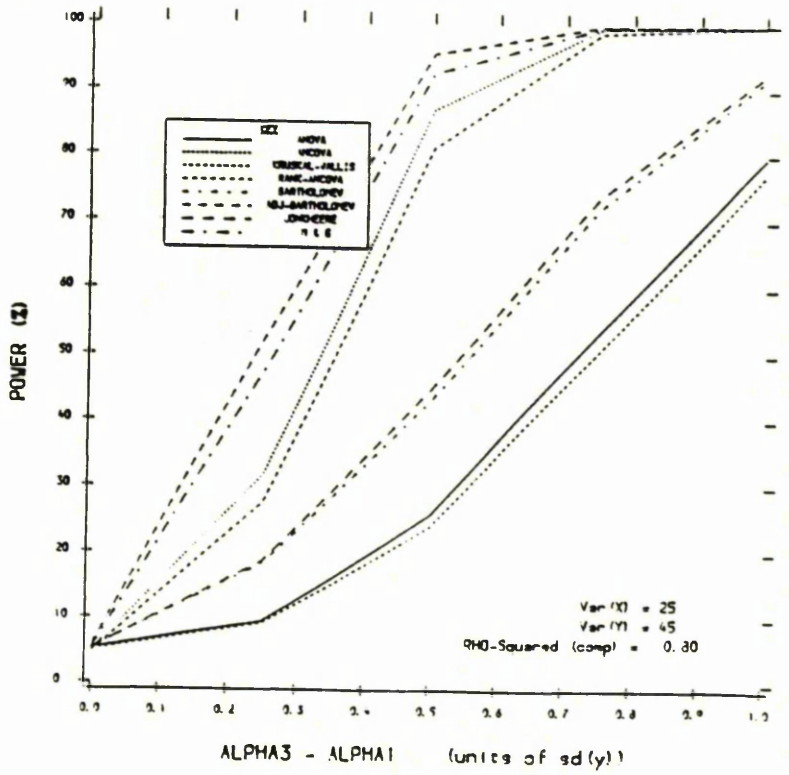


Figure 6.52 : $\rho^2_{comp} = 0$, Lognormal , $\sigma = 0.4$

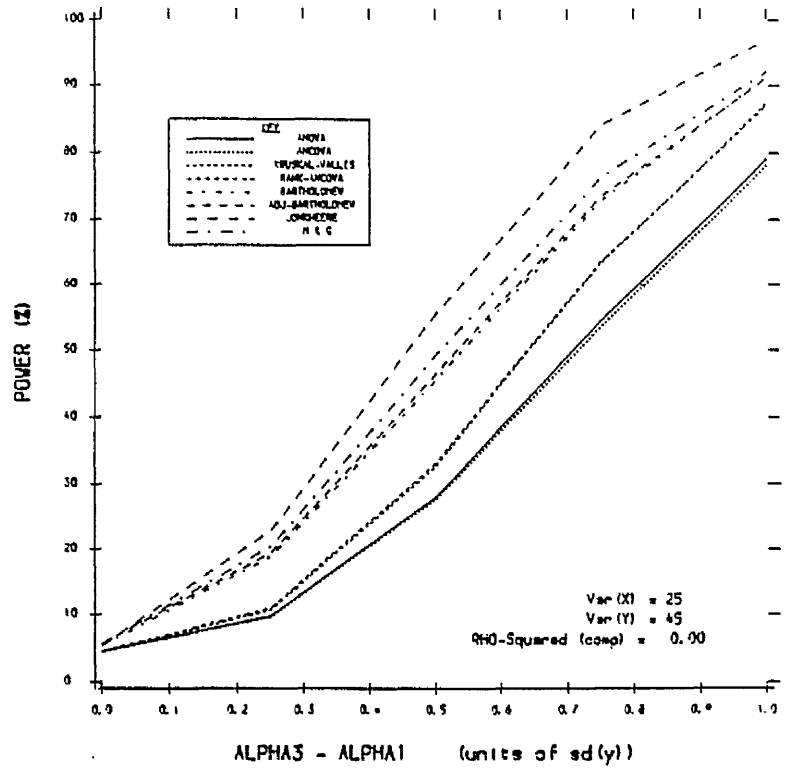


Figure 6.53 : $\rho^2_{comp} = 0.80$, Lognormal , $\sigma = 0.4$

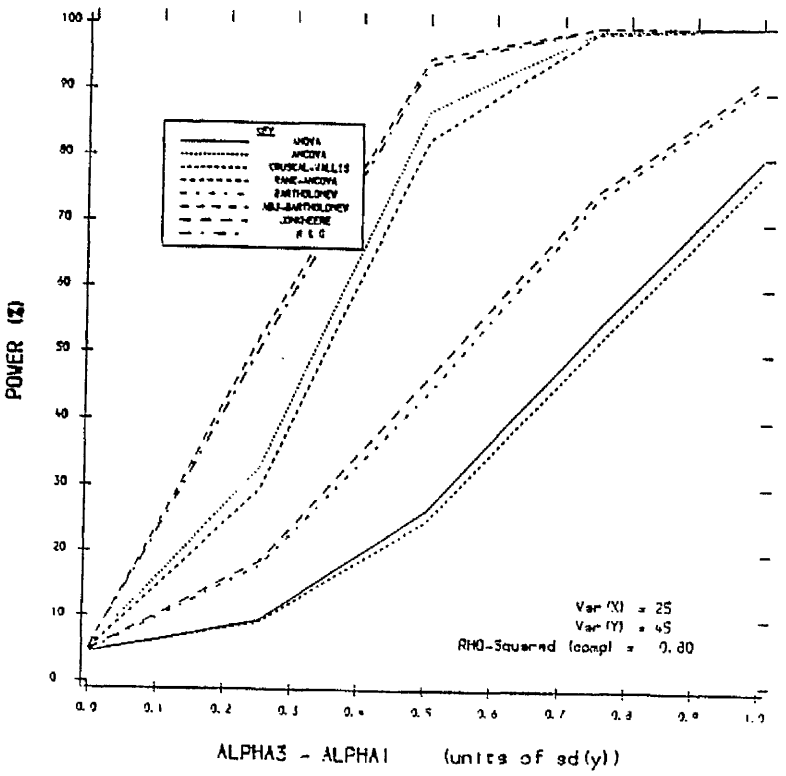


Figure 6.54 : $\rho^2_{comp} = 0$, Lognormal , $\sigma = 0.5$

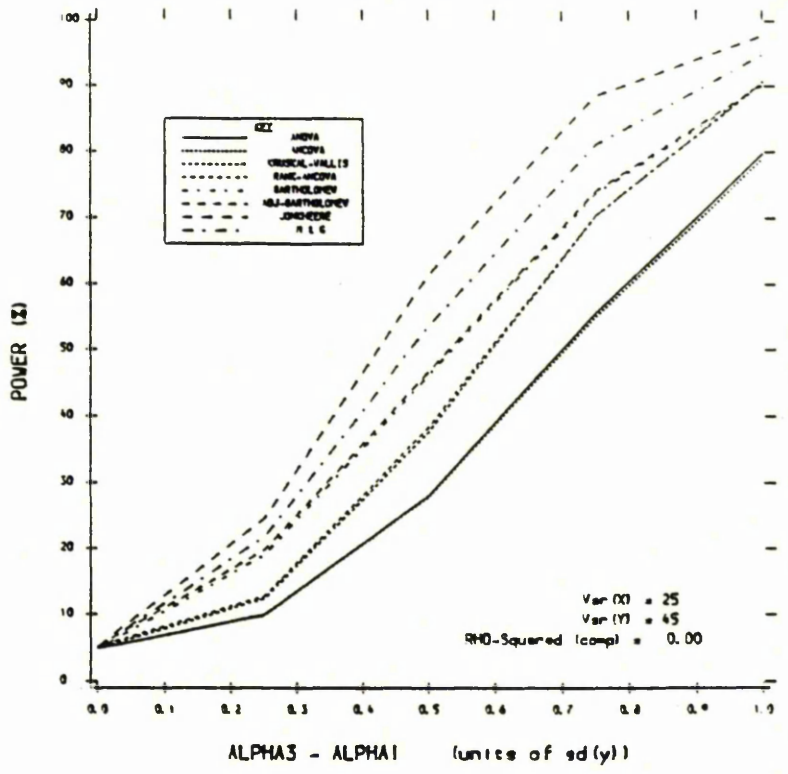
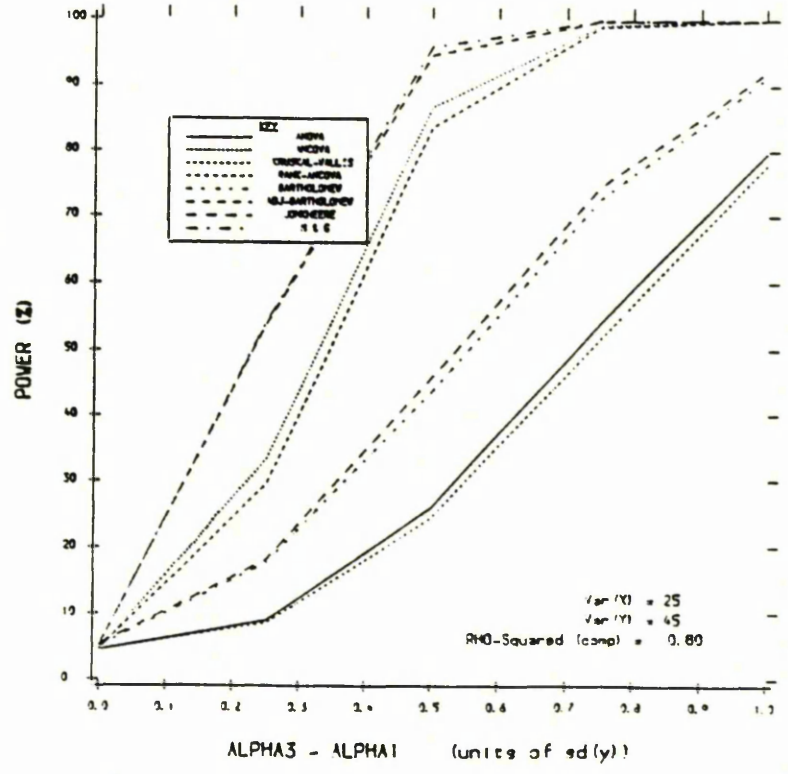


Figure 6.55 : $\rho^2_{comp} = 0.80$, Lognormal , $\sigma = 0.5$



the Normal errors situation was attributable to the poor performance of Test 8.

e.g Normal errors, zero correlation, and for both configurations of the group means : Test 8 performed very much worse than Test 6, where both tests incorporated ordering and (uninformative) covariates, but Test 6 had the additional assumption of Normal errors. The difference between these two tests was much greater than any other such pair of tests (i.e Tests 1 & 3 , Tests 2 & 4 and Tests 5 & 7).

For $\rho^2_{comp} = 0.2$, Test 6 still performed substantially better than Test 8. For the case of unequal spacing of the group means, the latter had similar power to Jonkheere's test (Test 7), and to Rank Analysis of Covariance (Test 4), despite incorporating more prior information. For $\rho^2_{comp} = 0.80$, Test 4 still out-performed Test 8 in certain circumstances.

An interesting feature which arose in connection with the Marcus and Genizi test was that other power curves commonly crossed that for the Marcus and Genizi test. The latter appeared to perform relatively well at small group separations but relatively poorly for larger separations. The gain in power as the group separation increased was relatively small compared with other tests.

It could be suggested that, perhaps, the poor performance of Marcus & Genizi's test could be attributed to a poor choice of k, the "smoothing parameter" in the matching function. This is unlikely, however, since the work of Section 6.2 would indicate that the performance of the test is relatively insensitive to the choice of k.

(ii) As observed in previous literature (e.g. Puri(1965)), Jonkheere's non-parametric test performed better than its Normal counterpart, Test 5, for equal spacing of the group means, but less well where all-except-one of the group means were equal.

(2) When the error distribution was Normal, the Normal tests performed better than their non-parametric counterparts, but not by a large margin.

For example, for a given correlation, and for the separation of the groups giving approximately 50% power for a given non-parametric test, the gain in power by using the corresponding Normal test was generally less than 5% . (This excludes

comparisons involving the Marcus and Genizi test, which performed unexpectedly poorly.)

(3) Only at the more extreme deviations from Normal errors did the non-parametric tests perform more favourably than their Normal counterparts.

(4) As expected, there was little, if any, difference between the results obtained in the presence of positively skewed errors and those obtained in the presence of negatively skewed errors.

The deviations in significance level from the nominal 5% were, in general, small enough to be attributable to random variability. (For 15000 simulations, the observed significance levels should be within 5 % ± 0.35 % . In only two cases were the observed significance levels outside this range.)

(5) As the error distribution became progressively non-Normal, the performance of the non-parametric tests increased, while the power of the Normal tests stayed approximately constant.

This was the pattern seen

(i) as β increased with γ held constant at $\gamma = 3$

(ii) as γ increased with β held constant at $\beta = 0$

(iii) as σ increased for the lognormal error distributions

(6) It is worth noting that the proposed test, $\bar{E}^2_{adj}$, performed very well, both in the presence of Normal errors and in the presence of non-Normal errors, for all values of correlation studied.

Normal Errors : In the presence of Normal errors, the proposed test, $\bar{E}^2_{adj}$ (Test 6) was the most powerful of all the tests compared, for $\rho^2 \gtrsim 0$, for both equal spacing of the group means and for unequal spacing. The margin of improvement of $\bar{E}^2_{adj}$ over $\bar{E}^2_{k}$ was much larger than the difference between any other such pairs of tests (i.e Tests 1 & 2 , Tests 3 & 4 and Tests 7 & 8). The test also performed favourably for $\rho^2 = 0$ (performing similarly to the best of the other tests).

Non-Normal Errors : Let ρ^2_{comp} represent the squared correlation

defining the Normal model to which a given set of results were to be compared. Then for $\rho^2_{comp} > 0$, in general $\bar{E}^2_{adj}$ still performed better than the other tests, with the exception of the cases with error distributions defined as :

(i) A mixture distribution with $\beta = 0$ and $\gamma = 5.4$

(ii) A lognormal distribution with $\beta = 1.75$ and $\gamma = 8.90$

where it performed marginally worse than its non-parametric competitor.

For $\rho^2_{comp} = 0$, in general, as would be expected, the non-parametric test intended for use in the absence of covariates (Test 7) performed best. $\bar{E}^2_{adj}$ generally performed second-best, followed by Test 8, Marcus and Genizi's test.

6.4 : Summary and Conclusions

The aim of Chapter 6 was to assess, using simulation, the benefits of inclusion of various forms of background information into the analysis of response data.

The background information was classified into three basic types :

(1) Covariate Information

(2) Information allowing a Normality assumption to to be made about the responses (or the responses given the covariates).

(3) Ordering information

The results obtained could be summarised as follows :

In general, the greater the amount of information which can be incorporated into an analysis, the more powerful the resulting test will be.

For example, even if a covariate is only moderately correlated with the response variable, its incorporation into an analysis can produce sufficient gains in sensitivity to make its inclusion worthwhile.

From the simulation results, it can be seen that for Normal errors, and for most cases using non-Normal errors which were considered, the tests assuming Normality performed better than their non-parametric counterparts. Even for the most severely non-Normal cases considered, any gains in power achieved by using non-parametric tests were minimal.

Thus, ideally, it would be recommended to use Normal-tests as a matter of course, and to incorporate information with respect to covariates and/or ordering where possible. One exception would be where it is desired to have as simple a test as possible. Then a non-parametric test would, generally, be more straightforward to apply than its Normal counterpart, and any losses of power incurred would not be large.

One practical problem which can arise is that when analyses are carried out on the basis of strong distributional assumptions about the data, this can arouse suspicion on the part of non-statisticians involved. Here, possibly the results of non-parametric analyses would be more credible, without much loss of statistical efficiency.

Chapter 7 : Practical Applications

7.1 : Introduction

The original aim of this chapter was to look at some published papers where order restricted inference would be appropriate, and to compare the published results to those which would be obtained by using the appropriate ordered inference methodology.

The sources of examples used were "The British Journal of Clinical Pharmacology" and "Clinical Pharmacology and Therapeutics" (years 1985-1989) which publish the results of clinical trials of all phases, as well as reporting advances in the pharmacological field.

As the tone of the first sentence of the chapter would suggest, this was not as straightforward as might be expected, for a variety of reasons :

(1) In the investigation of responses to differing drug doses, crossover studies were much more common than parallel group studies. It was not unusual to have as few as six individuals participating in studies of as many as five different dosing schedules.

(2) Often, there was insufficient information published even to reproduce the authors' own analyses, and certainly not enough to perform alternative analyses. Incidentally, all that is required in order to perform a Bartholomew's test is the sample means and their standard errors for the groups under study, plus, of course, the numbers of cases per group. However, often the standard errors were not supplied with the published results.

In the published papers, analyses tended to be based on either Analysis of Variance or Kruskal-Wallis tests, followed up by either tests or confidence intervals for differences between pairs of groups (often with no signs of correction for the multiple testing).

Despite there often being clear evidence of the appropriateness of order restricted inference, in no papers whatsoever were such testing procedures applied. This must, surely, be an indication of the need for better promotion of some

simple ordered testing procedures (e.g. Jonkheere's test would not be difficult to apply, in general, and has a simple approximate null distribution).

There now follow re-analyses of some studies for which there was sufficient published information to allow the application of alternative testing procedures. The poor standard of reporting of trials discussed above made it difficult to locate convincing examples. Indeed in order to find an example where raw data were available, the search needed to be extended beyond the more specialised journals. Thus the first example comes from a recent report in the British Medical Journal, whereas the remaining examples, all based on summary statistics, came from the journals set out at the start of this chapter.

7.2 : A Re-Analysis of the Plasminogen Activator Data

(See McNeill et al(1988))

The aim of this study was to investigate the thrombolytic efficacy of recombinant tissue plasminogen activator in acute myocardial infarction. This plasminogen activator works in a clot-specific manner (unlike more traditional treatments), activating the conversion of plasminogen to plasmin, which subsequently breaks down fibrin to fibrin degradation products.

In this study, fifty patients with acute myocardial infarction of four hours duration or less were randomised to receive either 20mg, 50mg or 100mg of recombinant tissue plasminogen activator intravenously, over ninety minutes.

After the ninety minute infusion period, the reperfusion grade of the affected coronary artery was assessed using coronary arteriography. Four grades of reperfusion were defined as follows:

(0) No perfusion
(1) Penetration with minimal perfusion
(2) Partial perfusion
(3) Complete perfusion .

In addition to the reperfusion data, various clotting factors were assessed after the ninety minute infusion period.

In the published paper, the reperfusion data were presented in the form of a 4 x 3 table of reperfusion grade against dose of plasminogen activator, and conclusions about the effects of the

activator on the reperfusion grade were drawn on the basis of a Chi-Squared test on that 4 x 3 table. Clearly such an analysis exploits neither the knowledge about the ordering of the doses of the activator nor the knowledge about the ordering of the reperfusion grades. In the following section, the results of a more appropriate analysis will be presented.

In the paper, for each of the clotting factor variables, the data were assessed using Analysis of Variance. Again, this does not make use of the prior ordering information available. A re-analysis of one such variable follows later. It should be noted that for some of the clotting factor variables presented in the paper, quite apart from being inefficient, the Analysis of Variance procedure would not even be appropriate, due to the widely differing variabilities in the groups under study. For example, for the total degradation products, the estimated standard deviations were 1.08, 4.55 and 11.35. One assumes that for such data, at the very least some form of transformation should have been applied before any analyses were performed.

The Re-Analyses

(a) A Re-Analysis of the Reperfusion Data

The reperfusion data were as shown below :

Reperfusion	**Dose of Tissue Plasminogen Activator**			
Grade	**20mg**	**50mg**	**100mg**	**Total**
0	4	4	1	9
1	4	1	2	7
2	3	5	4	12
3	5	7	10	22
Total	16	17	17	50

Due to the known thrombolytic properties of the tissue plasminogen activator, it would be natural to assume that as the dose of the compound increased, the thrombolytic effects would also increase i.e. the reperfusion grade would become higher. To re-analyse these data, a variant of Jonkheere's non-parametric

test making allowance for tied values was used, as shown below.

$$\text{Test Statistic : } \quad J = 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} p_{ij} - \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_i n_j$$

$$\text{where} \quad p_{ij} = \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} p_{irjs}$$

$$\text{with} \quad p_{irjs} = \begin{cases} 1 & \text{if } Y_{ir} < Y_{js} \\ \frac{1}{2} & \text{if } Y_{ir} = Y_{js} \\ 0 & \text{if } Y_{ir} > Y_{js} \end{cases}$$

When applied to the reperfusion data, a value of 205 was obtained for J, with an estimated variance of $^{227154}/_{18}$. This gave rise to a tail probablility of 0.034.

The conclusions from this analysis are in contrast with those which were published, where the three doses were not found to be significantly different with respect to the resultant perfusion grade (χ^2 = 6.22, on six degrees of freedom, leading to a tail probability of 0.40). This illustrates how the appropriate use of an ordered test can increase the sensitivity for detecting specified forms of group differences. It must be stated, however, that the analysis used in the paper was extremely insensitive. Some form of rank correlation test, or even a Kruskal-Wallis test on the data would have been preferable to the Chi-Squared test which was used. If a Kruskal-Wallis test, adjusted for ties, is performed, a value of 3.81 is obtained, which on reference to a Chi-Squared distribution on two degrees of freedom leads to a tail probability of 0.149. A better indication of the effect of incorporating ordering information is given by comparing the result of the Jonkheere's test (p = 0.034) to that from the Kruskal-Wallis test (p = 0.149), rather than comparing to the published Chi-Squared test result (p = 0.40).

(b) <u>A Re-Analysis of the Percentage Fibrinogen Data</u>

Due to the known mode of action of the tissue plasminogen

activator, it would be expected that as the dose increased, the percentage of fibrinogen remaining after the ninety minute infusion period would decrease. However, as mentioned earlier, no use was made of such information in the published analysis. These data will now be re-analysed using a Bartholomew's test.

Let $\hat{\mu}_i$ be the sample mean for the i^{th} group ($i = 1,2,3$) and let sd_i be the sample standard deviation for that group.

Let $\hat{\mu}_i^*$ be the isotonic estimate for the i^{th} mean. The results of Bartholomew's test could then be calculated as shown below.

	20mg	50mg	100mg
	i=1	i=2	i=3
n_i	16	17	17
$\hat{\mu}_i$	86.3	75.1	62.8
sd_i	14.92	14.12	20.61
$\hat{\mu}_i^*$	86.3	75.1	62.8

$$\hat{\mu} = \text{Grand Mean} = {}^{3725.10}/_{50} = 74.502$$

$$\text{Denominator of } \bar{E}_k^2 = \sum_{i=1}^{3} \{(n_i-1)sd_i^2 + n_i\hat{\mu}_i^2\} - N\hat{\mu}^2$$

$$= 17886.53$$

$$\text{Numerator of } \bar{E}_k^2 = \sum_{i=1}^{3} n_i (\hat{\mu}_i^* - \hat{\mu})^2 = 4561.0898$$

Thus $\bar{E}_k^2 = 0.255$.

This result is statistically significant, with tail probability of 0.00037. This compares with a tail probability of 0.001 if an Analysis of Variance test is used. Although it is clear that the overall conclusions of the two tests are the same, the strength of evidence available from the more appropriate Bartholomew's test is substantially greater.

7.3 : A Re-Analysis of the Cefotaxime Study

(See Matzke et al(1985))

Cefotaxime is a cephalosporin, effective in the treatment of most clinically relevant gram-positive and gram-negative bacteria. The aim of this study was to assess the kinetics of cefotaxime and its active metabolite, desacteyl cefotaxime in patients with varying degrees of renal function.

Subjects entering the study were classified into five groups, namely :

(I) Normal creatinine clearance ($CL_{cr} \gtrsim 90$ml/min)

(II) Mild renal insufficiency ($30 \leqq CL_{cr} \leqq 89$ ml/min)

(III) Moderate renal insufficiency ($16 \leqq CL_{cr} \leqq 29$ ml/min)

(IV) Severe renal insufficiency ($4 \leqq CL_{cr} \leqq 15$ ml/min)

(V) End-stage renal disease requiring maintenance haemodialysis ($CL_{cr} \lesssim 6$ ml/min)

Each subject received a single 1mg dose of cefotaxime, after which blood samples were drawn at fixed times and urine samples were collected, in order to determine the individual patient kinetics using established evaluation techniques.

In the published paper, separate Analysis of Variance procedures were used to compare each of the kinetic parameters among the different groups. Here, as an example, the results for C_{max}, the peak serum concentration for the desacetyl metabolite will be re-analysed.

Note that here, instead of the treatments applied having some prior ordering, it is the stage of advancement of the patients' disease state in the different groups which has a natural ordering. It could be expected that, since the drug clearance from the kidneys would be impaired in the groups with renal insufficiency, that the serum drug concentrations would therefore be higher in these groups, the magnitude of the concentration dependent on the degree of impairment. It would seem to make sense incorporate the prior ordering information, and employ some form of ordered testing technique on these data.

From the paper, it is not clear why the CL_{cr} data were categorised to produce the five patient groups. Rather than group the CL_{cr} data, possibly some form of rank correlation or robust regression approach would have been more appropriate.

The estimated means and standard deviations for the C_{max} parameter in the different groups are shown in the following section.

The Re-Analysis

Let $\hat{\mu}_i$ be the sample mean C_{max} value for the i^{th} group ($i = 1,2, \ldots, 5$) and let sd_i be the sample standard deviation. for that group.

Let $\hat{\mu}_i^*$ be the isotonic estimate for the i^{th} mean. The results of Bartholomew's test could then be calculated as follows :

	I	II	III	IV	V
	i=1	i=2	i=3	i=4	i=5
n_i	8	8	8	8	8
$\hat{\mu}_i$	7.9	9.4	8.8	14.7	21.0
sd_i	6.0	6.1	2.5	5.5	4.1
$\hat{\mu}_i^*$	7.9	9.1	9.1	14.7	21.0

$$\hat{\mu} = \text{Grand Mean} = {}^{494.4}/_{40} = 12.36$$

$$\text{Denominator of } \bar{E}_k^2 = \sum_{i=1}^{5} \{(n_i-1)sd_i^2 + n_i\hat{\mu}_i^2\} - N\hat{\mu}^2$$

$$= 1857.256$$

$$\text{Numerator of } \bar{E}_k^2 = \sum_{i=1}^{5} n_i (\hat{\mu}_i^* - \hat{\mu})^2 = 970.176$$

Thus $\bar{E}_k^2 = 0.522$.

(Significant, with tail probability of $p = 1.1*10^{-6}$).

Comparison of the Results of Bartholomew's Test to the Results of Analysis of Variance

For this example, clearly the method of analysis should be relatively unimportant, due to the large magnitude of difference between the groups. Even procedures with relatively poor sensitivity could pick up such overwhelmingly large group

differences easily.

As in the previous example, it was found that although the conclusions which could be drawn from the Analysis of Variance and the Bartholomew's tests were the same, the strength of evidence of group differences available from the more appropriate Bartholomew's test was greater (but both tests gave results which were highly statistically significant).

While a classical Analysis of Variance yielded a tail probability of $2.3*10^{-5}$, the result for a Bartholomew's test was even more significant with a tail probability of $1.1*10^{-6}$.

7.4 : A Re-Analysis of the Tolrestat Study

(See Raskin et al(1985))

In this study, the effects of various doses of the inhibitory compound, tolrestat, on the red blood cell sorbitol level were compared. Individuals entering the study were assigned at random to receive either

(I) A placebo

(II) Tolrestat, 25mg b.i.d.

or (III) Tolrestat, 100mg b.i.d.

After two weeks of treatment, the red blood cell sorbitol level was measured, and the percentage change from the pre-treatment baseline value was calculated.

In the published paper, conclusions were drawn on the basis of an Analysis of Variance procedure performed on these percentage changes, plus follow-up multiple comparisons. The conclusions were that while both of the doses of active compound were significantly different from placebo, there was no significant difference between effects of the two active doses.

Although the result of the Analysis of Variance was not given in the paper, the results will be calculated here. It must, of course, be borne in mind that in working purely with the published figures, one is always limited by the precision to which the published figures are given.

In these data, due to the nature of the compound being studied, and due to the type of measurements being analysed, there is a natural ordering of the expected magnitudes of the results in the different groups, in that the higher the dose of the inhibitor, the greater the expected magnitude of sorbitol

reduction would be. The data will now be analysed using both a Bartholomew's test and an Analysis of Variance test.

The Analysis

Let $\hat{\mu}_i$ be the sample mean difference for the i^{th} group ($i = 1,2,3$) and let se_i be its standard error. Let $\hat{\mu}_i^*$ be the isotonic estimate for the i^{th} mean. The results of Bartholomew's test were then calculated as follows :

	Placebo	25mg	100mg
	i=1	i=2	i=3
n_i	8	8	7
$\hat{\mu}_i$	20.4	-25.8	-57.5
se_i	13.8	11.3	13.2
$\hat{\mu}_i^*$	20.4	-25.8	-57.5

$$\hat{\mu} = \text{Grand Mean} = {}^{-445.7}/_{23} = 19.38$$

$$\text{Denominator of } \bar{E}_k^2 = \sum_{i=1}^{3} \{n_i(n_i-1)se_i^2 + n_i\hat{\mu}_i^2\} - N\hat{\mu}^2$$

$$= 48294.6$$

$$\text{Numerator of } \bar{E}_k^2 = \sum_{i=1}^{3} n_i(\hat{\mu}_i^* - \hat{\mu})^2 = 23161.3$$

Thus $\bar{E}_k^2 = 0.480$.

(Statistically significant, with tail probability of 0.00037).

When a standard Analysis of Variance procedure was performed on the same data, the test statistic obtained had an associated tail probability of 0.0015. Thus, as with the previous example, although the conclusions which could be drawn from the two tests are the same, the strength of evidence from the Bartholomew's test is greater.

7.5 : A Re-Analysis of the Felodipine Study

(See The Canadian Felodipine Study Group(1988))

The aim of this study was to look at the anti-hypertensive effects and side-effects of the calcium-antagonist felodipine in patients with persistent hypertension despite β-blocker therapy.

After a four-week placebo run-in period, patients with supine diastolic blood pressure exceeding 95mmHg were randomised to one of the following three treatment groups :

(1) Placebo

(2) Felodipine, 5mg per day

(3) Felodipine, 10mg per day ,

where their progress was monitored after two weeks and after four weeks of study treatment.

In the published paper, analysis of the changes in blood pressure was performed using multiple regression on drug dose, initial blood pressure and age. Whether the use of dose as a covariate was appropriate is doubtful, since there are usually problems of non-linearity of the dose-response relationship. Probably, what would be more appropriate would be to perform some form of ordered test, possibly after adjustment for the initial blood pressure and age. Unfortunately, the authors omitted to publish the means and standard errors for the blood pressure differences, so that even the simplest of ordered tests on the raw differences could not be performed.

The adverse effects associated with felodipine tend to be mainly flushing, plus side-effects related to fluid retention, for example ankle swelling and weight gain. Presented in the paper were the average weight gains and ankle swellings plus their standard errors for the different groups. Here the results of ankle swelling will be re-analysed. It could be expected that as the dose of felodipine increased, that the problem of ankle swelling would also increase, and so order restricted testing should be appropriate.

The means and standard errors for the changes in right ankle circumference at weeks two and four are shown in the following section.

In the paper, the analyses used were a global Analysis of Variance test followed up be pairwise t-tests between the groups (with no indication given of any adjustment of the t-deviates

used for the multiple testing).

The conclusions drawn from these analyses were that there was a "slight increase" in the ankle circumference for both felodipine groups after two weeks (compared to placebo), that there was a "significant increase" after four weeks, and that there were "no changes" in the placebo group.

The Re-Analysis

(a) Analysis of the Changes After Two Weeks

Let $\hat{\mu}_i$ be the sample mean difference for the i^{th} group ($i = 1,2,3$) and let se_i be its standard error. Let $\hat{\mu}_i^*$ be the isotonic estimate for the i^{th} mean. The results of Bartholomew's test were then calculated as follows :

	Placebo	5mg	10mg
	i=1	i=2	i=3
n_i	32	36	34
$\hat{\mu}_i$	-0.1	0.3	0.4
se_i	0.2	0.2	0.2
$\hat{\mu}_i^*$	-0.1	0.3	0.4

$$\hat{\mu} = \text{Grand Mean} = {}^{21.2}/_{102} = 0.208$$

$$\text{Denominator of } \bar{E}_k^2 = \sum_{i=1}^{3} \{n_i(n_i-1)se_i^2 + n_i\hat{\mu}_i^2\} - N\hat{\mu}^2$$

$$= 139.554$$

$$\text{Numerator of } \bar{E}_k^2 = \sum_{i=1}^{3} n_i (\hat{\mu}_i^* - \hat{\mu})^2 = 4.594$$

$$\text{Thus } \bar{E}_k^2 = 0.033$$

(Non-significant with tail probability of $p = 0.066$).

(b) Analysis of the Changes After Four Weeks

Using the same notation as in (a), the results of Bartholomew's test were as follows :

	Placebo	5mg	10mg
	i=1	i=2	i=3
n_i	24	31	29
$\hat{\mu}_i$	-0.1	0.6	0.8
se_i	0.2	0.2	0.2
$\hat{\mu}^*_i$	-0.1	0.6	0.8

$$\hat{\mu} = \text{Grand Mean} = {}^{39.4}/_{84} = 0.469$$

$$\text{Denominator of } \bar{E}_k^2 = \sum_{i=1}^{3} \{n_i(n_i-1)se_i^2 + n_i\hat{\mu}_i^2\} - N\hat{\mu}^2$$

$$= 103.240$$

$$\text{Numerator of } \bar{E}_k^2 = \sum_{i=1}^{3} n_i (\hat{\mu}^*_i - \hat{\mu})^2 = 11.480$$

Thus $\bar{E}_k^2 = 0.111$

(Statistically significant with tail probability of p = 0.002).

Comparison of the Results of Bartholomew's Test to the Results of Analysis of Variance

As in the previous examples, it was found that although the conclusions which could be drawn from the two types of test were broadly similar, the strength of evidence of group differences available from the more appropriate Bartholomew's test was substantially greater than that available from a classical Analysis of Variance.

For the Week 2 data, a classical Analysis of Variance yielded a tail probability of 0.191, compared to the much more "borderline" value of 0.066 for the Bartholomew's test. Similarly, for the Week 4 data, an Analysis of Variance yielded a tail probability of 0.008, compared with a value of 0.002 for the Bartholomew's test.

7.6 : Conclusions

The re-analyses of the published results illustrate how, with the use of appropriate order restricted inference, one is able to improve the statistical sensitivity for detecting specified forms of differences between treatment groups with an a priori ordering. For most of the examples considered, no major changes in conclusions resulted from using the appropriate methodology. However, for the example of Section 7.2, the conclusions drawn about the comparative effects of the applied treatments depended on the method of analysis used. While the published Chi-Squared test was fairly convincingly non-significant, the more sensitive Jonkheere's test produced a result indicating significant differences among the groups. It is not difficult to imagine cases where a non-significant, but borderline, result from an Analysis of Variance could correspond to a highly statistically significant result from an order restricted test. In that case, the conclusions drawn, and hence the plans for future research, could be different dependent on the method of analysis used.

The re-analyses shown in this chapter have worked purely on the basis of performing a single global test of equality versus some ordered alternative. What is more common is that if the global test was rejected, some form of follow-up procedure would be used to establish where the significant treatment differences lay. In the context of order restricted inference, a suitable follow-up closed testing procedure which preserves the required overall significance level was described by Marcus et al(1976).

Often, one would want to produce confidence intervals for the group means or for differences between groups means. For problems where order restricted inference is clearly appropriate, the confidence intervals could be based around the isotonic mean estimates, rather than around the original sample means.

Chapter 8 : Conclusions and Further Work

The aim of the work which has been described in the previous chapters was to look at two practical problems which are common in the context of clinical trials, but which are often dealt with inappropriately.

Chapters 2-4 looked at the problems of dealing with incomplete data while Chapters 5-7 looked at the potential benefits of incorporating ordering information into the analysis of response data.

In general terms, the techniques which are normally applied to these problems are not always invalid, but do always involve disregarding some of the useful information available.

In this thesis, various alternatives were put forward to replace the existing methods used in these problems, and to allow incorporation of more of the information available. Also, compromise solutions were provided for use in circumstances where the performance of complex analyses would not be possible.

In the analysis of incomplete multivariate data, the methods in common use (e.g. using only complete cases) tend to be invalid, and the presence of even relatively few incomplete cases can cause very large biases in the results obtained. Clinicians need to be made aware of this problem, and should be educated about the simple checks which can be made to assess whether there are likely to be severe bias problems in using the simpler methods. Statisticians, meanwhile, need to give more thought to the development of simple approaches which could be accepted by the clinicians without undue suspicion.

On the topic of incorporating ordering information, the picture is less clear. One can gain efficiency in analyses by using appropriate methodology, but this could be at the expense of losing credibility in the eyes of the clinicians involved. In effect, the appropriate methodology amounts to a generalisation of one-sided tests, and these can cause great controversy even in their simplest forms (see Salsburg(1989), Fleiss(1987), Fleiss(1989)). Further work in this area could run the risk of being an exercise in mathematical statistics, with little, if any, relevance to clinical medicine.

The basic overall conclusions were that, to achieve statistical efficiency, within the computational constraints in force, as much of the useful information available should be included as possible, be it ordering or covariate information, or information from incomplete cases, since the gains in precision and/or accuracy in the results obtained could be surprisingly large. However, the use of complex statistical procedures could prevent the results obtained being credible in the eyes of the clinicians involved.

As is always the case, the research described in this thesis could not possibly hope to answer all of the questions of interest in the areas covered. Rather, this work can only be thought of as a starting point and source of ideas for future research. On the topic of analysis of incomplete data, it would be interesting to look at some models for multivariate data where the covariance matrix is restricted to have a given structural form, for example, the compound symmetry model which is commonly used in the analysis of complete repeated measures data. Although some work has been done on fitting restricted models in the single-group case (see Jennrich and Schluchter(1986)), little such work has been done for the multiple-group problem.
(Some coverage is given in Murray(1989)).

On the topic of ordered-restricted inference, there are many possibilities for further research. For example, a common design of Phase II studies is where each individual receives in turn each of the treatments under investigation e.g. a single patient may receive the placebo, and the low dose and the high dose of the active compound, in a random order. Clearly the observations from a given case could not be assumed to be unrelated, and so some form of multivariate analysis would be required, for example by fitting a Multivariate Normal model to each case, then performing an overall test for ordering of the mean vector components. Although some early work is available for testing such multivariate hypotheses (see Kudo(1963)), little work seems to have been done in the intervening period. It would be interesting to look at such models with a view to incorporating covariate information, perhaps in some form of multivariate regression model with ordered intercept parameters. As a side issue, it could also be worth investigating the reasons for the poor performance of the Marcus and Genizi test used in the

simulations of Chapter 6.

Although the ideas for further research work are useful, probably the most important "further work", and, arguably, the most difficult to achieve, is the problem of how to gain acceptance of fairly complex statistical procedures in the medical community without losing credibility. It could be argued that it is far more important to heighten the awareness of clinical investigators to, for example, the possible problems of bias in the standard ad hoc analyses of incomplete data, and to suggest alternative analyses, rather than to advance research into more and more complicated statistical procedures, leaving the clinical parties involved behind.

Appendix 1 : Isotonic Regression and the Up and Down Blocks Algorithm

Definitions : (i) Let $X = \{ x_1, x_2, \ldots, x_k \}$, with the elements of X conforming to the simple order $x_1 \leq x_2 \leq \ldots \leq x_k$.

A function f on X is isotonic with respect to this ordering if

$$f(x_1) \leq f(x_2) \leq \ldots \leq f(x_k)$$

(ii) Let w(x) denote a positive weight function defined for x, and let g denote a given function of X.

Then, a function g^* is defined as an isotonic regression on g with weights w if and only if g^* is an isotonic function which minimises

$$\sum_{x \in X} (g(x) - f(x))^2 \, w(x)$$ among all such functions, f.

The Problem of Interest :

We have k samples drawn from Normal populations.

Let Y_{ij} = Observation j from Sample i

$(i = 1,\ldots,k \; ; \; j = 1,\ldots,n_i)$

with $Y_{ij} \sim N(\mu_i , \sigma^2)$, where it is known that the k means conform to an ordering $\mu_1 \leq \mu_2 \leq \mu_3$.

Let the sample means for the k groups be represented by $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$. It is of interest to produce maximum likelihood estimates for the population means which follow the known ordering conditions. Due to sampling variability, however, the sample means may not follow the desired ordering.

For the k samples, the likelihood function is given by

$$\frac{1}{\{ \sigma \sqrt{(2\Pi)} \}^N} \exp \left[-\frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(Y_{ij} - \mu_i)^2}{\sigma^2} \right]$$

where N is the total number of observations.

To maximise the likelihood, subject to the defined constraints, it can be seen that this is equivalent to finding values for $\mu_1, \ldots, \mu_k$ to minimise

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$$ subject to $\mu_1 \leq \ldots \leq \mu_k$

This can easily be shown to be equivalent to minimising

$$\sum_{i=1}^{k} n_i (\bar{Y}_i - \mu_i)^2 \qquad \text{subject to } \mu_1 \leq \ldots \leq \mu_k .$$

The Up and Down Blocks Algorithm

The identical nature of this problem and the isotonic regression problem defined in (ii) above can be seen by inspection.

One algorithm for performing isotonic regression, the 'Up and Down Blocks Algorithm' was described succinctly by means of a flow-chart in the 1972 book by Barlow et al(1972) (reproduced below).

The aim was to take a set of values, $X = \{x_1, \ldots, x_k\}$, and provide an isotonic regression for these. For the problem defined in the previous section, each $x_i \in X$ would be a sample mean.

For the algorithm, a block is defined as a set of consecutive values of X. The method begins with each individual x_i as a block on its own, and progressively pools together consecutive blocks if they do not follow the required ordering, continuing to pool until there are no further ordering violations. The flow-chart describing the operation of the algorithm was as shown below.

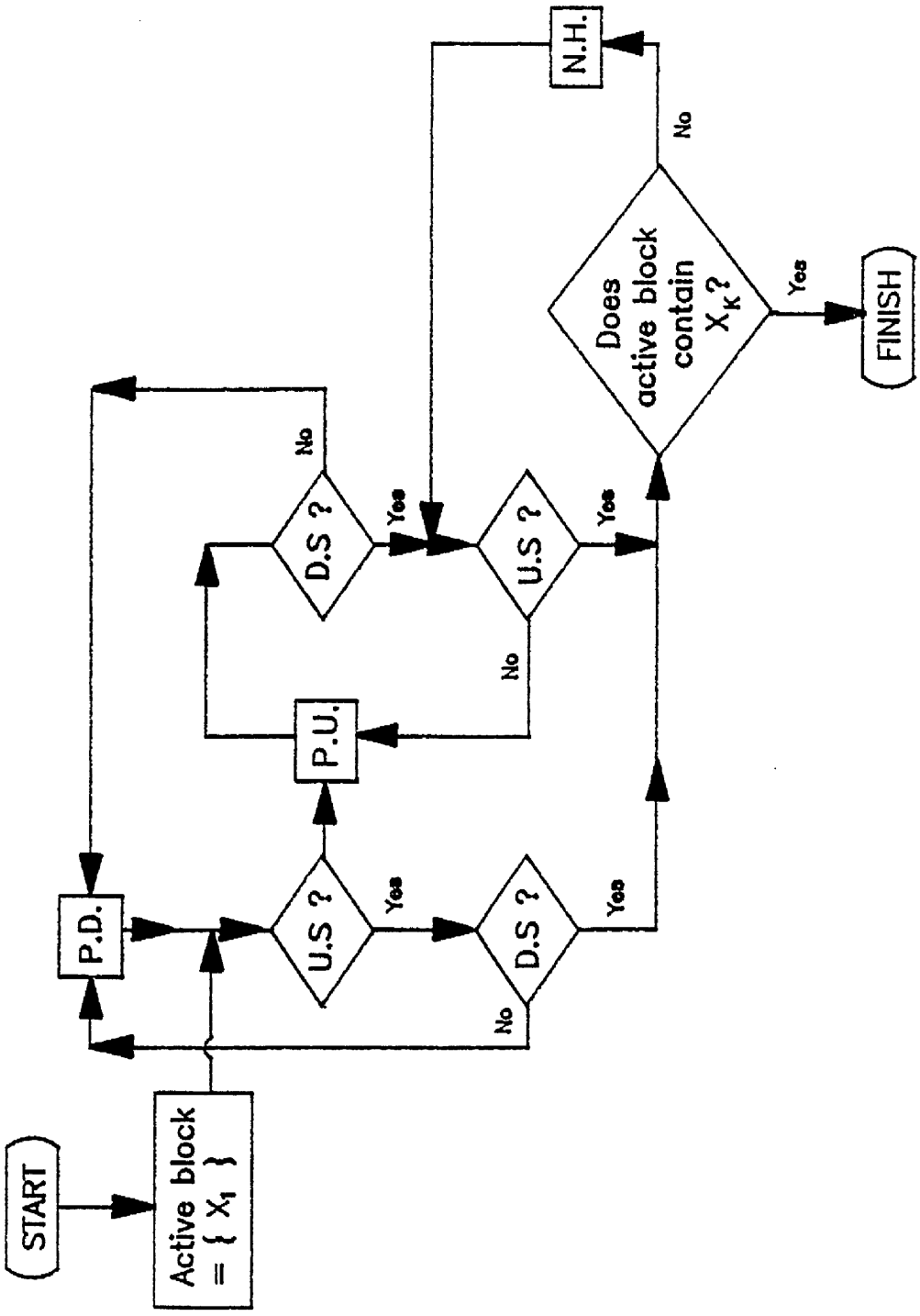
At each stage of the algorithm, one block is specified as active (i.e specified as a possible candidate for amalgamation with its surrounding blocks).

A block is defined as up-satisfied if the mean for that block is less than the mean for the immediately-following block. Similarly, a block is defined as down-satisfied if the mean for that block is greater than the mean for the immediately-preceding block.

In the chart, the abbreviations are :

U.S.	:	The active block is up-satisfied
D.S.	:	The active block is down-satisfied
P.U.	:	Pool the active block with the next higher block, the new block becoming active
P.D.	:	Pool the active block with the next lower block, the new block becoming active
N.H.	:	The next higher block becomes active

Figure A1.1 : The Up and Down Blocks Algorithm



REFERENCES

AIYAR R.J., GUILLIER C.L., ALBERS W. (1979)
Asymptotic Relative Efficiencies of Rank Tests for Trend Alternatives
Journal of the American Statistical Association 74 , 226-231

ANDERSON T.W. (1957)
Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations are Missing
Journal of the American Statistical Association 52 , 200-203

ARMITAGE P., McPHERSON K., ROWE B.C. (1969)
Repeated Significance Tests on Accumulating Data
Journal of the Royal Statistical Society Series A 132 , 235-244

BARLOW R.E., BARTHOLOMEW D.J., BREMNER J.M., BRUNK H.D. (1972)
Statistical Inference Under Order Restrictions
(Wiley, London)

BARNARD G.A. (1963)
Discussion on Professor Bartlett's Paper
Journal of the Royal Statistical Society Series B 25 , 294

BARTHOLOMEW D.J. (1959)
A Test of Homogeneity for Ordered Alternatives
Biometrika 46 , 33-48

BARTHOLOMEW D.J. (1959)
A Test of Homogeneity for Ordered Alternatives II
Biometrika 46 , 328-335

BARTHOLOMEW D.J. (1961)
Ordered Tests in the Analysis of Variance
Biometrika 48 , 325-332

BEALE E.M.L., LITTLE R.J.A. (1975)
Missing Values in Multivariate Analysis
Journal of the Royal Statistical Society
Series B 37 , 129-145

BOISSEL J.P., BLANCHARD J., PANAK E., PEYRIEUX J.-C., SACKS H. (1989)
Considerations for the Meta-Analysis of Randomized Clinical Trails
Controlled Clinical Trials 10 , 254-281

THE CANADIAN FELODIPINE STUDY GROUP (1988)
Antihypertensive Efficacy of the Calcium-Antagonist Felodipine in Patients with Persisting Hypertension on β-adrenoceptor blocker therapy
British Journal of Clinical Pharmacology 26 , 535-545

CHACKO V.J. (1963)
Testing Homogeneity Against Ordered Alternatives
The Annals of Mathematical Statistics 34 , 945-956

CHALMERS T.C., SMITH H. Jr, BLACKBURN B., SILVERMAN B., SCHROEDER B., REITMAN D., AMBROZ A. (1981)
A Method for Assessing the Quality of a Randomised Clinical Trial
Controlled Clinical Trials 2 , 31-49

COX D.R., HINKLEY D.V. (1974)
Theoretical Statistics
(Chapman and Hall, London)

CREPEAU H., KOZIOL J., REID N., YUH Y. (1985)
Analysis of Incomplete Multivariate Data from Repeated Measures Experiments
Biometrics 41 , 505-514

DEMPSTER A.P., LAIRD N.M., RUBIN D.B. (1977)
Maximum Likelihood From Incomplete Data via the EM Algorithm
Journal of the Royal Statistical Society Series B 39 , 1-38

DOKSUM K. (1967)
Robust Procedures for Some Linear Models With One Observation per Cell
The Annals of Mathematical Statistics 38 , 878-883

EFRON B. (1971)
Forcing a Sequential Experiment to be Balanced
Biometrika 58 , 403-417

FLEISS J.L. (1987)
Letter to the Editor : Some Thoughts on Two-Tailed Tests
Controlled Clinical Trials 8 , 394

FLEISS J.L. (1989)
Letter to the Editor : One-Tailed Versus Two-Tailed Tests : Rebuttal
Controlled Clinical Trials 10 , 227-230

GNANADESIKAN R. (1977)
Methods for Statistical Data Analysis of Multivariate Observations
(Wiley , New York)

GOULD A.L. (1980)
A New Approach to the Analysis of Clinical Drug Trials With Withdrawals
Biometrics 36 , 721-727

GRIZZLE J.E. (1965)
The Two-Period Change-Over Design and its Use in Clinical Trials
Biometrics 21 , 467-480

HAMMERSLEY J.M., HANDSCOMB D.C. (1964)
Monte Carlo Methods
(Wiley, New York)

HARTLEY H.O. (1958)
Maximum Likelihood Estimation from Incomplete Data
Biometrics 14 , 174-194

HARTLEY H.O., HOCKING R.R. (1971)
The Analysis of Incomplete Data
Biometrics 27 , 783-823

HERRICK A.L., WALLER P.C., BERKIN K.E., PRINGLE S.D., CALLENDER J.S., ROBERTSON M.P., FINDLAY J.G., MURRAY G.D., REID J.L., LORIMER A.R., WEIR R.J., CARMICHAEL H.A., ROBERTSON J.I.S., BALL S.G., McINNES G.T. (1989)
Comparison of Enalapril and Atenolol in Mild to Moderate Hypertension
The American Journal of Medicine 86 , 421-426

HILLS M., ARMITAGE P. (1979)
The Two Period Cross-Over Clinical Trial
British Journal of Clinical Pharmacology 8 , 7-20

HOCKING R.R., MARX D.L. (1979)
Estimation With Incomplete Data : An Improved Computational Method and the Analysis of Nested Data
Communications in Statistics - Theory and Methods A8 (12) 1155-1181

HOLLANDER M. (1967)
Rank Tests for Randomised Blocks When the Alternatives Have an a priori Ordering
The Annals of Mathematical Statistics 38 , 867-877

HUITEMA B.E. (1980)
The Analysis of Covariance and Alternatives
(Wiley, New York)

JENNRICH R.I., SCHLUCHTER M.D. (1986)
Unbalanced Repeated-Measures Models With Structured Covariance Matrices
Biometrics 42 , 805-820

JOHNSON N.I., KOTZ S. (1969)
Distributions in Statistics
Continuous Univariate Distributions - 1
(Houghton Mifflin, Boston)

JONES B., KENWARD M.G. (1989)
Design and Analysis of Crossover Trials
(Chapman and Hall, London)

JONKHEERE A.R. (1954)
A Distribution-Free K-Sample Test Against Ordered Alternatives
Biometrika 41 , 133-145

KERSHNER R.P., FEDERER W.T. (1981)
Two Treatment Crossover Designs for Estimating a Variety of Effects
Journal of the American Statistical Association
76 , 612-618

KRUSKAL J.B. (1964)
Nonmetric Multidimensional Scaling : A Numerical Method
Psychometrika 29 , 115-129

KUDO A. (1963)
A Multivariate Analogue of the One-Sided Test
Biometrika 50 , 403-418

LAIRD N., LANGE N., STRAM D. (1987)
Maximum Likelihood Computations With Repeated Measures : Application of the EM Algorithm
Journal of the American Statistical Association
82 , 97-105

LEHMANN E.L. (1964)
Asymptotic Nonparametric Inference in Some Linear Models With One Observation per Cell
The Annals of Mathematical Statistics 35 , 726-734

LEWIS J.A. (1989)
Discussion on Paper by Murray and Findlay
Statistics in Medicine 8 , 1302-1303

LITTLE R.J.A. (1979)
Maximum Likelihood Inference for Multiple Regression With Missing Values : A Simulation Study
Journal of the Royal Statistical Society Series B 41 , 76-87

LITTLE R.J.A. (1988)
Robust Estimation of the Mean and Covariance Matrix from Data With Missing Values
Applied Statistics 37 , 23-38

LITTLE R.J.A., RUBIN D.B. (1987)
Statistical Analysis With Missing Data
(Wiley, New York)

LITTLE R.J.A., SMITH P.J. (1987)
Editing and Imputation for Quantitative Survey Data
Journal of the American Statistical Association 82 , 58-68

MACHIN D., CAMPBELL M.J. (1987)
Statistical Tables for the Design of Clinical Trials
(Blackwell Scientific Publications, Oxford)

MAGEL R. (1983)
A Comparison Between the Jonkheere and Chacko Nonparametric Test Statistics When Sample Sizes are Small
Technical Report,
Department of Mathematics,
North Dakota State University.

MANN H.B., WHITNEY D.R. (1947)
On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other
The Annals of Mathematical Statistics 18 , 50-60

MARCUS R., GENIZI A. (1987)
Nonparametric Analysis of Covariance With Ordered Alternatives
Journal of the Royal Statistical Society Series B 49 , 102-111

MARCUS R., PERITZ E., GABRIEL K.R. (1976)
On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance
Biometrika 63 , 655-660

MARDIA K.V., KENT J.T., BIBBY J.M. (1979)
Multivariate Analysis
(Academic Press, London)

MARRIOTT F.H.C. (1979)
Barnard's Monte Carlo Tests : How Many Simulations ?
Applied Statistics 28 , 75-77

MATTHAI A. (1951)
Estimation of Parameters from Incomplete Data With Application to Design of Sample Surveys
Sankhya 2 , 145-152

MATZKE G.R., ABRAHAM P.A., HALSTENSON C.E., KEANE W.F. (1985)
Cefotaxime and Desacetyl Cefotaxime Kinetics in Renal Impairment
Clinical Pharmacology and Therapeutics 38 , 31-36

MENIERT C.L. (1986)
Clinical Trials : Design, Conduct, and Analysis
(Oxford University Press, Oxford)

McINNES G.T., MURRAY G.D. (1988)
Clinical Trials : Aims and Methods
Medicine International , 2447-2450

McNEILL A.J., SHANNON J.S., CUNNINGHAM S.R., FLANNERY D.J., CAMPBELL N.P.S., KHAN M.M., PATTERSON G.C., WEBB S.W., ADGEY A.A. (1988)
A Randomised Dose Ranging Study of Recombinant Tissue Plasminogen Activator in Acute Myocardial Infarction
The British Medical Journal 296 , 1768-1771

MORREY G.H. (1985)
A Simple Alternative to the Two-Period Crossover Design
Leicester Statistics Technical Report, No.2

MURRAY G.D. (1990)
Meta Analysis
British Journal of Surgery 77 , 243-244

MURRAY G.D., FINDLAY J.G. (1988)
Correcting for the Bias Caused by Drop-outs in Hypertension Trials
Statistics in Medicine 7 , 941-946

MURRAY L.S. (1989)
Modelling the Recovery After Severe Head Injury
Ph.D. Thesis
University of Glasgow

NAIR V.N. (1986)
On Testing Ordered Alternatives in Analysis of Variance Models
Biometrika 73 , 493-499

POCOCK S.J. (1978)
The Size of Cancer Clinical Trials and Stopping Rules
British Journal of Cancer 38 , 757-766

POCOCK S.J. (1983)
Clinical Trials : A Practical Approach
(Wiley, London)

PURI M.L. (1965)
Some Distribution-Free K-Sample Rank Tests of Homogeneity Against Ordered Alternatives
Communications in Pure and Applied Mathematics 18 , 51-63

QUADE D. (1967)
Rank Analysis of Covariance
Journal of the American Statistical Association 62 , 1187-1200

QUADE D. (1974)
Nonparametric Partial Correlation
in "Measurement in the Social Sciences - Theories and Strategies" 369-398
(Blalock H.M. Jr. ed)
(Aldine, Chicago)

QUADE D. (1982)
Nonparametric Analysis of Covariance by Matching
Biometrics 38 , 597-611

RANDLES R.H., WOLFE D.A. (1979)
Introduction to the Theory of Nonparametric Statistics
(Wiley, London)

RASKIN P., ROSENSTOCK J., CHALLIS P., RYDER S., MULLANE J.F., GONZALEZ R., HICKS D., SMITH T., DVORNIK D.(1985)
Effect of Tolrestat on Red Blood Cell Sorbitol Levels in Patients with Diabetes
Clinical Pharmacology and Therapeutics 38 , 625-630

ROBERTSON T., WRIGHT F.T., DYKSTRA R.L. (1988)
Order Restricted Statistical Inference
(Wiley, New York)

ROSENDORFF C., MURRAY G.D. (1986)
Ketanserin versus Metoprolol and Hydrochlorothiazide in Essential Hypertension : Only Ketanserin's Hypotensive Effect is Age Related
Journal of Hypertension 4 , S109-S111

RUBIN D.B. (1974)
Characterizing the Estimation of Parameters in Incomplete Data Problems
Journal of the American Statistical Association 69 , 467-474

RUBIN D.B. (1976)
Inference and Missing Data
Biometrika 63 , 581-592

SALSBURG D. (1989)
Use of Restricted Significance Tests in Clinical Trials : Beyond the One- Versus Two-Tailed Controversy
Controlled Clinical Trials 10 , 71-82

SCHWARTZ D., FLAMANT R., LELLOUCH J. (1980)
Clinical Trials
(Academic Press, London)

SEN P.K. (1960)
On Some Convergence Properties of U-Statistics
Calcutta Statistical Association Bulletin 10 , 1-18

SHAPIRO S.H., LOUIS T.A. (1983)
Clinical Trials : Issues and Approaches
(Marcel Dekker, New York)

SHORACK G.R. (1967)
Testing Against Ordered Alternatives in Model I Analysis of Variance ; Normal Theory and Nonparametric
The Annals of Mathematical Statistics 38 , 1740-1753

SILVERMAN B.W. (1986)
Density Estimation for Statistics and Data Analysis
(Chapman and Hall, London)

TERPSTRA T.J. (1952)
The Asymptotic Normality and Consistency of Kendall's Test Against Trend When Ties are Present in One Ranking
Indagationes Mathema 14 , 327-333

TITTERINGTON D.M., SMITH A.F.M., MAKOV U.E. (1985)
Statistical Analysis of Finite Mixture Distributions
(Wiley, Chichester)

YUSUF S., PETO R., LEWIS J., COLLINS R., SLEIGHT P. (1985)
Beta Blockade During and After Myocardial Infarction : An Overview of the Randomised Trials
Progress in Cardiovascular Disease 27 , 335-371

ZELEN M. (1979)
A New Design for Randomized Clinical Trials
New England Journal of Medicine 300 , 1242-1245