

THE UNIVERSITY *of* EDINBURGH

Thesis scanned from best copy available: may contain faint or blurred text, and/or cropped or missing pages.

Scanned as part of the PhD Thesis Digitisation project
http://libraryblogs.is.ed.ac.uk/phddigitisation

EVALUATION AND QUALITY CONTROL IN THE HAEMATOLOGY LABORATORY

Ewan W. MacArthur

Doctor of Philosophy

University of Edinburgh
1978

EX DIBL. UNIV. EDINBURGH

Acknowledgements.

I would like to thank my supervisors for their help and encouragement while preparing and writing this thesis. Dr. S.H. Davies, Consultant Haematologist, Royal Infirmary Edinburgh, gladly gave his time and experience to discuss clinical and haematological problems. His knowledge, enthusiasm and clarity of answer, was greatly appreciated, not only in the understanding of problems directly relating to this thesis, but also in many other areas. Similarly, Dr. R.J. Prescott, Medical Computing and Statistics Unit, University of Edinburgh, gave sustained help from the inception to the final form of the thesis. His advice in statistical matters and in the interfacing of statistics and medicine, before and during work on this thesis, has been a moulding force for my understanding and practice in medical statistics.

I would like to acknowledge the help given to me by Mr. P. Newman, Senior Chief Technician, Department of Haematology, Royal Infirmary Edinburgh. Discussions with him concerning general laboratory practice, quality control and the Coulter Counter Model S provided a rational and practical basis for much of my knowledge of clinical laboratory practice.

I would like to thank Mr. W. Lutz, Director, Medical Computing and Statistics Unit, University of Edinburgh, for his support in the preparation of the questionnaire and for the financial assistance made for computing.

I would like to thank the organisers of the DHSS (LDAG) and BCSH Haematology Quality Control Trials, Dr. S.M. Lewis and Mr. P.G. Ward, Royal Postgraduate Medical School, London, for their agreement to surveying participants of this trial, and also their help in the distribution of the questionnaires.

Finally, I would like to thank Elaine Livsey, Lynne Westwood and Tove Korstvedt for typing this thesis.

Abstract.

The major historical developments in clinical measurement and also in statistics and computing are surveyed, and an outline of the problems to be considered in the thesis is given.

Before dealing with clinical measuring processes in general, some methods for handling measurements are examined. In particular, procedures for dealing with replicate measurements are investigated.

The evaluation of a single clinical measuring process is discussed with special reference to continuous auto-analysers such as the Coulter Cell Counter Model S. The main facets of evaluation of a single clinical measuring device are presented. Designs and the corresponding analyses are given for the simultaneous estimation of the accuracy, precision and carryover in the presence of drift. In some situations, linearity is also investigated. The results of some practical experiments are presented.

To complete the discussion of the evaluation of measuring processes, the comparison of different clinical measuring processes for the same parameter(s) is investigated. A number of different methods of analysis are examined. Using clinical data, the results obtained from these analyses are presented.

In a survey of quality control the basic requirements for control in the clinical field relevant to haematology, and in the haematological laboratory, are discussed. Quality control is seen to have two components - Intra and Inter laboratory control. An examination of both types of control is presented. These incorporate discussion of methods and procedures of control and related topics. An example of an intra-laboratory scheme, currently in use, is presented. In the discussion of inter-laboratory control, the results of a survey of the participants of the Department of Health and Social Security and British

Committee for Standardisation in Haematology. Haematology Quality Control Trials are presented.

Finally conclusions, further areas for investigation and areas of development are presented and outlined.

Contents

Chapter 1. Introduction

1.1 Historical Developments 1

1.2 Outline of Thesis 16

Chapter 2. Statistics and Measurement Handling

2.1 Introduction 19

2.2 Statistical Background 19

2.2.1 Statistical Notation 19

2.2.2 Statistics in Practice 21

2.3 Handling Numerical Results 28

2.3.1 The Effect of Rounding 28

2.3.2 Measuring by Matching 33

2.3.3 Outliers 36

2.3.4 Estimation of Clinical Parameters in the Presence of Outliers 39

2.3.5 Common Procedures for Estimation in the Presence of Outliers 44

2.3.6 Results of Simulation of a Measuring Process and Estimating Procedures 47

Chapter 3. Evaluation of a Machine/Method

3.1 Introduction 63

3.2 Description of a Clinical Measuring Process 63

3.3 Physical Evaluation 65

3.4 Description of Coulter Counter Model S and Model S Senior 68

3.5 Definitions of Statistical Terms Concerning Evaluation 72

3.5.1 Accuracy 72

3.5.2 Precision 73

3.5.3 Carryover 74

3.5.4 Drift 79

3.5.5 Linearity 80

3.6 Comments on Previous Investigations 81

3.7 Proposed General Model 84

3.8 Design Considerations 88

3.9 The Use of Coding 94

3.10 Estimation of Parameters 98

3.11 Investigations of Triple and Double Designs for Uncodable Situations 106

3.12 Analysis of Practical Experiments 108

3.13 Construction of Designs: The General Situation 130

3.13.1 Introduction 130

3.13.2 Codable Levels: The General Situation 131

3.13.3 Designs Using Three or More Levels 140

3.13.4 Uncodable Levels: The General Situation 144

3.14 Conclusions 145

Chapter 4. Comparison of Machines/Methods

4.1 Introduction 147

4.2 Methods and Indices of Comparisons 150

4.2.1 Diagnostic Correlation 150

4.2.2 Traditional Measures of Association 152

4.2.3 General rXc Tables 152

4.2.4 Reliability Models 157

4.2.5 Numerical Comparisons of Indices 158

4.2.6 Loss Function Approach 160

4.2.7 The Use of Loss Approach 162

4.3 Correlation 164

4.4 Regression 170

4.5 Specific Methods of Comparing Numerical Results 175

4.5.1 Genesis of the Problem 175

4.5.2 Problems Encountered in Estimation 178

4.5.3 Practical Solutions to Estimation Problem 182

4.5.4 Numerical Example 185

4.6 Two Further Specific Solutions 194

4.6.1 Introduction 195

4.6.2 Outline of the Basic Models: Grubbs, Replicate 196

4.6.3 Estimation 198

4.6.4 Comparison of Estimators from Models 204

4.6.5 Estimation of Bias 208

4.6.6 Hybrid Schemes 211

4.6.7 Hybrid Scheme for Two Machines 213

4.6.8 Numerical Example 215

4.6.9 Maximum Likelihood 220

4.6.10 Tests and Related Problems 224

4.6.11 Conclusions 232

4.7 Instrumental Variables 233

4.8 Non-Equivalent Machines/Methods 235

4.9 Conclusions 241

Chapter 5. Intra-Laboratory Quality Control

5.1 Introduction 245

5.2 Normal Ranges (Reference Ranges) 246

5.2.1 Introduction 246

5.2.2 Construction of Normal Ranges: Parametric 251
5.2.3 Construction of Normal Ranges: Non-Parametric 262
5.2.4 Discussion 266
5.3 Monitoring a Laboratory's Clinical Measurements 268
5.3.1 Introduction 268
5.3.2 Standards and Reference Materials 271
5.3.3 Quality Control and Frequency of Monitoring 273
5.4 Control Charts 278
5.4.1 Introduction 278
5.4.2 Shewhart Charts (Standard Deviation Plots) 278
5.4.3 Cusum Control Charts 290
5.5 Estimation of the Process Mean 295
5.5.1 Introduction 295
5.5.2 Proposed Estimators 297
5.5.3 Comments on the Estimators of the Population Mean 303
5.6 Formal Quality Control Schemes using Patient Data 305
5.6.1 Introduction 305
5.6.2 The Number Plus Method 306
5.6.3 The Method of 'Average of Normals' 310
5.6.4 A Modal Control Procedure 312
5.6.5 Conclusions 315

5.7 Quality Control for Multiple Determinations 316

5.8 Outline of a Practical Scheme 321

5.8.1 Introduction 321

5.8.2 Reference Preparations 326

5.8.3 The Daily Set-Up 330

5.8.4 Through the Day Control 332

5.8.5 Comments on the Scheme 345

Chapter 6. Inter-Laboratory Quality Control

6.1 Introduction 350

6.2 Aims of Inter-Laboratory Schemes 351

6.3 Set-up of Inter-laboratory Quality Control Schemes 355

6.4 True Value, Assigned True Value and Reference Laboratories 356

6.5 A General Scheme for Inter-Laboratory Comparison 362

6.6 The Youden Plot and Modifications 368

6.7 Scheme of Zender and Linder 373

6.8 The DHSS (LDAG) and BCSH Haematology Quality Control Trials 375

6.9 Survey of Participants of the DHSS (LDAG) and BCSH Haematology Quality Control Trial 380

Tables of Responses to Questionnaire DHSS Blood Trials 394

Chapter 7. Conclusions, Proposals and the Future 413

Appendix 1

Appendix 2

Appendix 3

Appendix 4

Appendix 5

Book References

References

1 Introduction.

1.1 Historical Developments.

Some of the problems of measurement facing the haematologist today can be tackled by the application of current statistical methods, or the development of new statistical techniques. As a basis for the understanding of such problems, it is necessary to review briefly the historical development which has led to them.

The origin of medicine has been investigated by many authors. Castiglioni (B 7) has sketched the following historical development. If we understand the term "medicine" to indicate the relief of pain by our own means or with the help of others, or the repair of damage produced by injury or disease, we should think first of all of its instinctive origin, much as the first expression of pain at instruction". Castiglioni suggests that there was a natural development through simple empirical medicine, 'magic' medicine to priestly medicine although the empiricism flourished throughout. By this stage, different civilizations had developed their own empirical art or 'magic/priestly' medicine. We cannot hope to trace ancient medicine, the methods and techniques used. However, we give a brief historical outline of some of the most important medical and technical-medical developments.

Hippocrates, 460-355 B.C., raised the 'art' to establish a theoretical framework and introduce the concept of an ethic. Further, he advocated, for example, the estimation of temperature, the investigation and 'measurement of arterial pulsations' and listening for noises in the chest. Certainly this was primitive measuring, but it was a further 2000 years before temperature was measured more accurately

clinically. The 'great dogmatist and omniscient physcian' Galen, 138-201 A.D., could be considered as terminating the period of ancient thought. His encyclopaedic writings were the basis for much of the medical practice up till the Renaissance. The Greeks, Arabs and Chinese did experiment but to a large extent such investigations were lost until the great intellectual upsurge of the Renaissance. Medicine, perhaps did not suffer such a great stagnation as did other fields of knowledge.

With the Renaissance came great developments in the 'exact sciences' and this can be seen as propagating developments in medicine. In fact, we see the birth of 'exact medicine', with the basis of observation and measurement. Leonardo Da Vinci contributed to medicine, among many other fields. He ignored the writings of Galen and preceded using observation, measurement and his undoubted genius. In the early 17th century, Harvey announced his observations on circulating blood, and within sixty years, Lower had transfused blood from dog to dog. Later he transfused lamb's blood to man. This was also done in France by Denic (Denys) whose patients unfortunately died. This was instrumental in a bill banning such transfusion procedures. (The dangers of transfusion were not 'scientifically' investigated till 1875 by Landois, (see Castiglioni (B 7)) although 347 transfusions had been recorded by then).

Red corpuscles were indentified in the 17th century. One form of white corpuscles was recognised by the 18th century. However, it was not until the 19th century that the leucocyte and its role in inflammation and disease control was discovered. Undoubtly this advance could not have taken place without an optical device and indeed without other techniques such as staining.

The use of optical devices and other instruments greatly helped

the evolving medicine. The evolution had taken medicine from a magical practice or at best the practice of procedures handed down from the ancients, to the frontiers of science. Lyonet, 1720, is credited with the first anatomical microscope, (see Shyrock (B33)), but by 1830, it has been suggested that the majority of the English text-books spoke of microscopy as an 'art now almost forsaken'.

The use of such devices, however, brought a sequence of rapid and important discoveries, starting with the discovery of leucocytes, phagocytosis, platelets and coagulation theories. The use of instruments which gained greater and greater precision was important in the study of the physiology of the circulation. During such investigations, Ludwig introduced methods of recording medical data graphically. The beginning of the 19th century was an important time in the development of bio-chemistry. After the studies of Liebig, 1803-73, (see Castiglioni (B 7), which brought about the study of what we now recognise as organic chemistry, developments in tests, techniques and laboratory methods helped and motivated clinical investigations and research. The use of spectrophotometry and the quantitive study of respiratory gases also proved useful.

The use of blood indices to reflect a patient's health status has a relatively short history. As we have indicated, blood analysis had been carried out by the ancient Greeks. The theory of 'humours' was based on the layers observed when blood was left standing. The humours were the erythrocyte layers, the leucocyte and platelet layer and the plasma layer. One of the earliest 'measures' was suggested by Fahraeus - the Erythrocyte Sedimentation Rate, ESR. Fahraeus (1918) had observed that the sedimentation of red cells was altered during pregnancy. Further, he later reported, Fahraeus (1921), that is was also accelerated in the presence of many diseases.

The variation in the physical characteristics of erythrocytes in disease had been known almost since the first days of the microscope study of blood. Vierordt, around 1860, recorded the earliest erythrocyte counts, (see Castiglioni (B 7)).

Since Hoppe-Seyler, in the 19th century had investigated the spectroscopy of haemoglobin and isolated crystals of haemoglobin and other blood pigments, establishing the chemical formulae for some, the study of haemoglobin and its importance had been accepted.

The main problem with the use of suggested blood indicies was the lack of accuracy and reproducibility of primary results. Further, the procedures to obtain these measures were complex. A number of indicies were used before 1929. For example. a well known index was the Colour Index. In 1903, Capps suggested the Volume Index (= Volume of a Red cell/Volume of a 'normal' red cell). Haden suggested the Saturation Index (= weight of Haemoglobin in cell/weight of Haemoglobin in a 'normal' cell). Haden (1923) examined the bahviour of some of the current blood indicies in the differentiation of anaemias. These were - red blood cell count, oxygen capacity, haemoglobin, percentage haemoglobin, haematocrit, Volume Index, Colour Index, Saturation Index, volume of the average cell, weight of haemoglobin in the average cell and the percentage of haemoglobin in the average cell. Jorgensen and Warburg (1927) carried out a similar review of the current blood indicies with special reference to their variability in pernicious anaemia.

Wintrobe (1929) presented new blood indicies, which were based on currently used indicies - the red cell count, the haemoglobin and haematocrit. These indicies were the mean cell volume, the mean cell haemoglobin and the mean cell haemoglobin content, which are used today. In a series of review papers. Wintrobe (1930, 1931-32,

1933a, 1933b), reviewed methods for determining these indicies, suggesting the best and the most reproducible methods for their estimation. He further investigated the 'normal' values attained for these indicies by healthy people. From that point in time, these indicies appeared to gain acceptance.

A major advance occured with the introduction, by Coulter in 1956, of a method of counting red cells by electrical resistance. This allowed greater precision and accuracy, and it has enabled many more specimens to be handled.

With the Renaissance a surge of quantification allowed physics and astronomy to develop laws and otherwise advance quite spectacularly. The same may have been hoped of medicine, but this was not so. The biological field found itself immersed in a genuinely complex system. Many different fields of knowledge had to expand before medical advancement. For example, the hermetical sealing of a Rey thermometer in 1644 by Ferdinand II of Tuscany brought about a 'modern' temperature measuring device. But, it was not a part of clinical hardware till the middle of the 1700's. Singer (B34), in his 'Short History of Medicine' has accounted for the long delay in the development of the instruments involved by the difficulties encountered. The beginning of the 18th century saw Floyer introduce his 'physician's pulse watch'. However, it does not appear to have been accepted by his colleagues for some time. Some thirty years later Hale introduced a method of measuring blood pressure. With better instruments and techniques, an evolving consistent theoretical framework and not least investigators with knowledge of other fields and a desire to link areas of endeavour, we arrive at the present day. This is at the root of the multidisciplinary approach which is gaining favour in medicine today.

The Renaissance had brought a new surge, namely to quantify.

However, by the 18th century, Shyrock (B33) has written -

'The reaction against the mechanical and mathematical schools involved a rejection of quantitative procedures as such. In the first flush of the new order all things seemed possible. Could not remedies and cures be devised, as well as instruments and machines? Would not research sooner or later prove of benefit to the actual practice of medicine?'.

Unfortunately it was much more difficult to reconcile observation, experiment and theory in the medical sphere as compared to say physics. Foucault (B15) has described the situation as follows.

'It is understandable, then, that medicine should turn away from what Sauvages called a 'mathematical' form of knowledge: 'Knowing quantities and being able to measure them, being able, for example, to determine the force and speed of the pulse, the degree of heat, the intensity of pain, the violence of the cough, and other such symptoms', Meckel measured, not to obtain knowledge of mathematical form, but to gauge the intensity of the pathological quality that constituted the disease. No measurable mechanics of the body can, in its physcial or mathematical particularities, account for a pathological phenomenon; convulsions may be due to dehydration and contraction of the nervous system - and this is certainly a phenomenon of a mechanical order; but it is a mechanics of inter-linked qualities, articulated movements, upheavals that are triggered off in a series, not a mechanics of quantifiable segments. It may involve a mechanism, but it cannot belong to the order of Mechanics as such'.

It would appear from what we have said above, that we not only require measuring instruments, but a theory and understanding of what significance these observations have to the clinical state of the patient. Thus medical theory does not only require a physiological

flowcharting of functions and a general anatomical plan, but requires an understanding of the relationship between measurement and the clinical status of the population or the patient. We need to know how measurements on the same parameters for different patients and for different clinical states behave. The obvious extension are such things as 'normal ranges'.

In France, particularly during the first part of the 19th century, there was a move to collect and investigate clinical measurements. Shyrock notes that 'the French school, which did so much to develop statistical medicine, did relatively little to invent or exploit new instruments'. We now trace the major developments in statistics and compare its development with that of medicine.

Although games of chance are found in ancient times, the use of probability in investigating chances of winning only became advanced after the Renaissance. Gambling appears to be one of the motivating forces in such study. With the increasing emphasis in science now on measurement, much more attention was paid to the nature of errors and the investigation of measurements in general. In 1713 James Bernoulli's work describing the binomial distribution was published, posthumously. Forty years later, De Moivre showed that the Normal (Gaussian) distribution was a limiting case of the binomial (see Yule and Kendall (B38)). In 1763, a cleric Thomas Bayes, published a treatise on probability (see Bartlett (B5)). The major theorem still bears his name and is employed for example in computer aided diagnosis. By the end of the 18th century Laplace and later Gauss had become interested in the nature of errors of measurement and their influence to the specific investigation. (Much of this work was done on astronomical data).

The first recorded use of 'statistics' was in preventative medicine, by Mather in 1721 (see Shyrock (B33)). He compared a group of inoculated

people with a group of untreated people during an epidemic. The collection of vital records and their analysis has a much older history than this. In fact, by 1657 published figures giving causes of death and death rates were regularly published. Graunt (1620-74) made some attempt to analyse these figures see Castiligoni (B 7). Basically these were the only attempt to use 'statistics' till much later. Louis, in the 19th century, used statistics in estimating the efficiency of different treatments. He was an enthusiastic worker in using such methods, but was roundly critisied by Bertillon because he did not take into account the probable error in his results.

The use of statistics in social studies and also medical problems was exhibited in 1869 by Quetelet (see Shyrock (B33)). We find that by the end of the century much more interest was shown in the use of statistics in experimental work. In many studies we find the use of correlation analysis. Pearson (B30) was instrumental in pointing out the problem with 'spurious correlation'. (As an example of this if we consider the annual figures of the number of horse-drawn carriages on the road and the number of maternal mortalities, we find a strong correlation. However, there is no casual relationship in the simultaneous reduction of these figures. In almost any body of data, such 'correlations' can be found).

From this period, through the work of Pearson, Gosset ('student') and Fisher, we find much of the now commonly accepted statistical methods and procedures and terms introduced. Both theoretical and practical developments are numerous. In 1924, Shewhart introduced the control chart as a method of recording and controlling industrial processes. Many of his ideas have only been slightly altered, and can be found in many laboratories today. As late as 1930, Neyman and Pearson gave a general (and perhaps the most widely accepted) general

theory of hypothesis testing (see Kendall and Stuart (B 23)).

The much used and important group of techniques called Analysis of Variance, were introduced in the mid-1930's. Much of the work was initiated by Fisher through his work at Rothamstead Experimental Station. Sequential methods were developed by Wald during the early 1940's. Much of his work fell under the American official secrets act and was not cleared till after the war ended. As a particular result of this work, the basis was laid for cumulative sum (cusum) techniques, which appeared about 1950, and have since found application in the quality control procedures used in clinical laboratories.

We have traced the development of both medicine and statistics. The actual number of clinical measurements has increased. These measurements are made for many different reasons. In some cases we wish to be able to classify from our measurements whether a patient is 'normal' or 'abnormal'. In other cases, we wish to monitor a patient's treatment scheme.

An increased awareness of statistical techniques in medicine has accompanied the proliferation of the number of clinical parameters measured and the frequency of their measurement. This proliferation has been caused by the development of rapid and accurate laboratory methods and electronic methods of handling measurements and information. The development of these electronic devices has, of course. a short history. It is largely dependent on the development of semi-conductors, transistors, micro-circuits, fibre-optics and other recent innovations. In the modern laboratory many of the clinical measurements can be determined in a very short time and from a very small amount of patient material. Today there are auto-analysers which can carry out seven blood count determinations on a patient sample simultaneously, processing many hundreds of patient samples in a working day. A

similar situation exists in clinical chemistry. The Technicon SMAC autoanalyser can return twenty to twenty four determinations on a single pa tient sample. Not only are the actual measurements of these machines dependent on electronic devices, but the actual control of the machine's complicated sequencing and self-monitoring systems are directed by the central 'computer' mechanism. Often the results can be directed to storage or analysis by a laboratory computer or to a central computer through a laboratory link.

The history of computers, as we would recognise them, is only probably at most a quarter of a century long, although calculating machines have been employed for much longer than this. Pascal in 1642 had demonstrated a machine which added. Forty years later Leibnitz produced a device which could also multiply. But the technology just did not exist to enable any recognisable advance. With only slight modifications, these machines have the same basic features exhibited by the mechanical machines found frequently until the last ten years. Babbage, 1792-1871, spent much time and effort investigating automatic calculating machines. His son, in 1906, built a machine based on this work. However, the first truly automatic calculating machine did not appear until 1944. It was slow and had very little storage. Eckert and Mauchly developed an electronic automatic calculator and demonstrated it in 1946. The first stored program machine appeared three years later. In 1950, the first computers became available on the open market. These machines were very large physically. A computer occupying a room then would fit on a desk top today, and be slower in operation than a modern pocket calculator. These first computers depended on valves and generated a vast amount of heat. The revolution came when transistors, micro-chips, micro-circuits, semi-conductors etc., became available, thus allowing

the devices to be reduced in size. This also allowed greater storage and flexibility. The end of this 'evolution' was the interest shown by the commercial world and research and development institutions in a robust, 'cheap' multi-purpose machine.

As we have mentioned previously, mechanisation in the clinical laboratory is a common feature today. One of the major effects of the increased mechanisation in medicine is the 'distance' between the clinician and the measuring of the patient sample. The measurement is often carried out by specially trained technicians. However, the automation of these standard clinical determinations does not do away with the basic requirement that the measurement procedures should be monitored.

The philosophy of measurement, in itself is a large and complex topic. It is not proposed that we should explore this topic in detail, although we should be aware of some of the basic concepts. Let us suppose that we wish to measure a certain haematological parameter e.g. haemoglobin concentration. In measuring the haemoglobin concentration of a specific patient sample, we require to assign a numerical value to the specific sample. Obviously the number must be meaningful to the clinician. For this to be so, the rules by which we assign numbers to patient samples must be the same for different samples i.e. the rules defining the measuring process are the same irrespective of when the measurement is made. (It should also not depend on the place where the measurement is made, but this will be discussed more fully later). Basically this implies that the measurements must come from a measuring process which is 'controlled'. We shall see that this objective is not trivially attained.

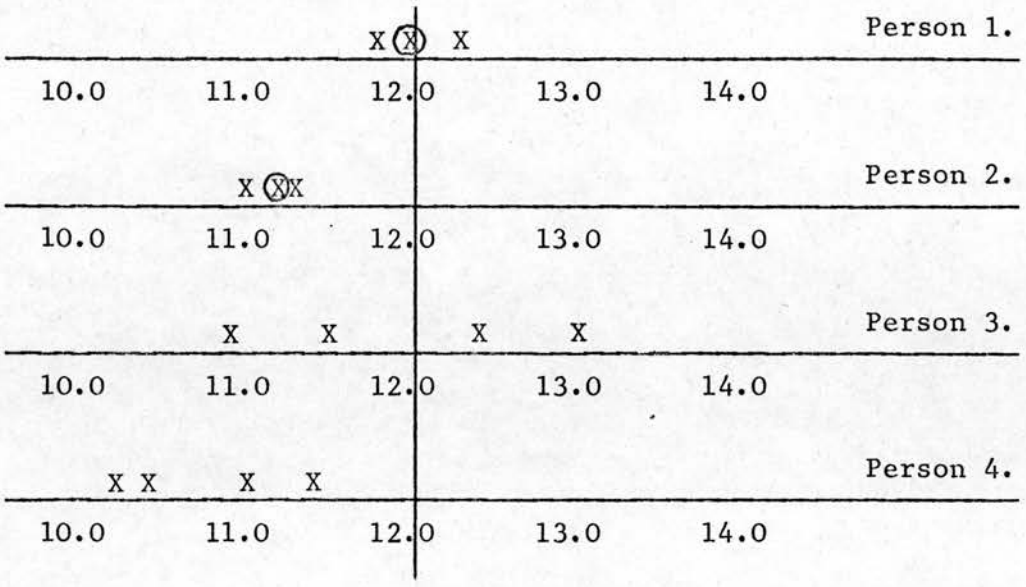
The clinician, having obtained measurements from a 'controlled' process can now attach some clinical meaning to the results. The

measurement process is 'controlled', but is this sufficient for the clinician's requirements? He surely must have measurements with a degree of sensitivity which at least meets his requirements. Let us explain this by again considering the estimation of haemoglobin concentration. Do we need to be able to measure this parameter correct to six decimal places? To answer this question, we need to know why the measurements are being made. If we simply wish to classify a patient as 'normal' or 'abnormal', we probably do not require to have such a finely divided scale as when we are monitoring the effects of a certain treatment in this parameter.

Two important characteristics of a measuring process are their accuracy and precision. To illustrate these concepts, let us consider the following example. We ask four different people to measure the haemoglobin concentration of a specific patient's speciment. This parameter has to be measured four times by each person. Now let us suppose that the true value of the haemoglobin concentration is 12.0 g/dl, this being unknown to any of the measurers. The results of these estimations, in g/dl, are -

Person 1.	12.1	12.0	11.9	12.0
2.	11.0	11.1	11.1	11.2
3.	11.0	11.5	12.5	13.0
4.	11.0	10.5	10.2	11.6

Let us plot these results as shown in figure 1.1.



X- Measurement

Ⓧ - Two coincident measurements

Person 1. can be described as accurate - on average his results are the same as the true value, and also precise - the repeat measurements are 'close' together.

Person 2. can be described as precise but not accurate.

Person 3. can be described as accurate but not precise.

Person 4. is neither accurate nor precise.

We have said that the person is or is not accurate and/or precise. This rather gives the impression that these characteristics are personalised. This is not totally accurate. The person carrying out the measurements we shall consider as part of the measuring process. It is really the measuring process which is described by these characteristics. We should bear in mind, that by accepting the above characteristics - accuracy and precision - which will be discussed in more detail later, we have defined two parameters of the process which must be measured. That is we require to measure the accuracy and the precision of that measuring process. This would appear to be

'chicken and egg' situation. We have a measuring process and to determine whether it is operating in control or not, we require to make a series of measurements on the actual process. We will show that this situation can be dealt with competently.

As we have noted, the measurement of a clinical parameter requires a 'controlled' measuring procedure. An important point made in the following discussion is that to have meaningful measurements, we must be clear as to what we mean by a measuring process. Let us consider once again the measurement of haemoglobin concentration. If we have a series of patient samples for measurement it would seem at first sight, that to measure them we simply record the results from our machine or measuring method. The results can then be accepted as our measurements on the patient samples. Is this meaningful? If we knew that our machine/method always gave the true value of any patient sample, with respect to haemoglobin concentration, it would be sufficient. However, this is not so. There exists documented evidence to show us that different methods of measuring haemoglobin will give different results on the same specimen. This could be due to a simple calibration error. On the other hand, it may be due to a fundamental difference in what is being measured. For example, in the estimation of haemoglobin, usual clinical methods employ colour or light-intensity matching. These methods measure inert pigments of haemoglobin as well as the active haemoglobin. Comparison of such results with results from methods in which the oxygen-combining capacity is measured, show them to be different by at least 2% (Dacie and Lewis (B 9)). Thus the clinical results will often depend on the measuring methods chosen. Further the use of different reagents can often produce results even with the same methods of measuring.

Each method will have its own degree of reproducibility i.e.

precision - repeated measurements on the same specimen do not produce exactly the same value. If we investigate more fully, we see that these factors are not the only ones affecting the variability of our clinical results. There are environmental factors. For example, the operating staff, humidity, climate, pollen-count etc. could perhaps, all affect our measurements. Further, the collection and handling of the specimens may be factors affecting our process. The specimens could have been collected by a local G.P., special clinic or student doctor. The techniques used and the materials may all influence our measurements.

We have identified many factors which may affect our measurements. The measuring process must be seen in the light of these possible influences. Further, we must also define what group of patients we can measure by the process. This means listing all types of people, bloods etc. which do not yield 'correct' results from the process. (For example, Widmark and Weschler (1970) have shown that the Technicon Auto-Analyser when used for platelet counting gives falsely high values when the specimen being measured has a high leucocyte count, nucleated erthyocytes, Howell-Jolly bodies or malaria parasites. Hence such specimens should not be measured by this method).

We have given brief outlines of the development of medicine, with special emphasis on medical measurements and the increasing use of statistics. This discussion has been extended to show that with the development and expansion of electronic computers, medical measurement has become more extensive and much more frequently requested. We have also seen the 'modern' measuring process in influenced by many factors. This leads us to investigate our meaning of 'controlled'. We shall leave further discussion of these points till later. At this stage we shall give a brief outline of the thesis.

1.2 Outline of Thesis.

In the following chapters we discuss, with relation to some of the most common haematological parameters, what we mean by a measuring process. Having discussed the components of a measuring process, the problems of assessing these processes and of controlling them can be investigated. We study these problems, explaining methods currently used and suggesting improvements where necessary. Practical solutions to problems investigated are proposed. That is, methods and procedures are given which are not just theoretically valid, but also procedures which can be used in the haematology laboratory.

The thesis has two parts. The first part deals with the evaluation of measuring methods and the comparision of different methods. The second covers the control of such measuring processes.

The evaluation of a method as we point out, has two major parts. The physical evaluation deals with the actual hardware, its ease of use, breakdown profile and servicing and its effect on the laboratory staff. The statistical evaluation deals with the analysis of the results on a representative group of specimens i.e. estimation of accuracy, precision, drift, carryover, etc and the comparision with similar estimates from other methods. Although the former is much more likely to be an absolute evaluation the latter will not usually be of this nature. We require to assess accuracy, precision etc. in the light of experience or more properly, to compare directly with other processes. To have an absolute evaluation would require us to compare the results of our measurements with the 'true' values. Even in clinical chemistry, where a method proposes to measure say, sodium concentration in sera, the comparis on of machine results with 'known' standards, although perhaps the closest to absolute evaluation, does

not allow for the effects of some of the factors we have previously identified. In haematology, difficulties are experienced because of the rapid biological deterioration, and the special requirements of storage and handling. There are also problems in obtaining a true aliquot on which peripheral blood counts can be made. Further, there is a problem which is discussed later, concerning standards and reference preparations.

Although absolute evaluation of a method is not possible, we can investigate the accuracy, precision carryover etc. of a particular process. These may be considered to be the particular innate characteristics of the particular process. The evaluation of reagents is also important. This evaluation can be seen as assessing the stability of the reagent and/or assessing the compatibility of the reagent with the method. The evaluation of a method in complete isolation is of limited interest. Evaluation, even the evaluation of the innate characteristics of a process can have no meaning unless we compare the results with the results from an existing standard or reference method. Clinically it is of interest to compare methods not only to gauge precision etc., but basically to ascertain whether patients will be classified as having the same clinical status by different methods. The problems of comparison are discussed with special reference to existing 'approximate' solutions. It is shown that these methods may result in misleading conclusions and we suggest more appropriate solutions.

Such deliberations, however, concerning the evaluation and/or the comparison of methods must be done in the understanding that the process(es) involved are 'in control'. Fundamentally, it should be recognised that a process which is not in control produces results which have no meaning clinically or otherwise. Control has two components -

a process must be in control as judged against itself and possible standard or reference materials, but also it must be in control as judged against other processes purporting to measure the same clinical parameters (e.g. other laboratories). These two components of control are given the names intra (within) and inter (between) laboratory quality control. We discuss, methods of intra-laboratory control, and also methods of inter-laboratory control. The relationship between these facets of control are discussed. Finally, we present results of a survey carried out among the participants of the DHSS(LDAG) and BCSH Haematology Quality Control trials.

Finally, the results of these investigations are drawn together to give conclusions which have a sound theoretical basis and are also practical. Resulting from these conclusions, certain proposals are made concerning evaluation and quality control. Having carried out a detailed investigation of some of the problems found in a routine haematology laboratory, it is hoped that a direction and motivation for future research has been given.

Chapter 2. Statistics and Measurement Handling

2.1 Introduction

In this chapter, we introduce the basic statistical notation which will be used in the thesis. The notation is, as far as possible the standard notation which is found in the majority of statistical text-books. Perhaps the only immediate difference is that we shall refer to a Normal random variable as a Gaussian random variable. This is a valid and not uncommon alternative name. The main reason for this is to avoid confusion in future chapters between the clinical condition 'normal' and the type of random variable.

In this chapter we also discuss, briefly, some ideas in the estimation of statistical parameters. In the discussion we emphasise the necessity to be clear about the assumption which can be made concerning the statistical model being used, and how, if the assumptions change, certain accepted estimators may not necessarily be optimal.

Following these general comments we discuss some specific examples in which the handling of data can affect the final inferences which are made. In particular, for example, we investigate estimation of clinical parameters in the presence of 'outliers'.

2.2 Statistical Background

2.2.1 Statistical Notation.

We introduce the statistical notation, which will be used in this and the remaining chapters. The notation used is similar to that used in most standard text-books.

Let X be a continuous random variable with probability density function (p.d.f.), f(x) $-\infty<x<\infty$. The probability that the random variable takes a value between a and b is defined as follows.

$$P(a<X<b) = \int_a^b f(x)dx$$

The expected value of a random variable X, with p.d.f., f(x) is defined,

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

The expected value, E(X) is the theoretical or population mean of the random variable. We can define the population variance in a similar manner.

$$V(X) = E((X-E(X))^2)$$

In the specific case when X is a Gaussian (Normal) random variable, we have

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\tfrac{1}{2}(\tfrac{x-\mu}{\sigma})^2\} \quad -\infty<\mu<\infty,\ \sigma>0,\ -\infty<x<\infty$$

Further it can be shown that

$$E(X) = \mu \qquad V(X) = \sigma^2$$

A concise notation for the statement, 'X is a Gaussian (Normal) random variable with mean μ and variance σ^2', is

$$X \sim N(\mu, \sigma^2)$$

In the particular case when X has a Standard Gaussian Distribution, we have

$$X \sim N(0,1)$$

In this particular situation we denote, f(x) by $\phi(x)$. The cumulative function, $\Phi(x)$, often allows statements to be written more succinctly. For the standard Gaussian distribution we have

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\} \quad -\infty<x<\infty$$

$$P(X < x) = \Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$

When investigating two, or more random variables the joint probability density function is required. This is a natural extension of the univariate case. For two variates X and Y, we have the joint p.d.f. f(x,y). The expected value and variance of the variate are defined in a similar way to the above. A further quantity, the covariance of X and Y is defined to be

$$\text{Cov}\ (X,Y) = E((X-E(X))(Y-E(Y)))$$

2.2.2 Statistics in Practice

In laboratory measuring procedures, as in many other situations, it is common to make several measurements on the same specimen. Each measurement is assumed to be an estimate of the true, but unknown value of the parameter of interest. Usually this is done by splitting the specimen into several parts - aliquots, and measuring each aliquot under similar conditions with respect to the parameter of interest. In common usage, we make replicate measurements on the specimen. This, however, is not the same as in the 'physical' situation. Suppose we make replicate measurements of length or voltage across a piece of metal wire. In each of these cases the measurement is made on exactly the same material. When we split our specimen into aliquots, we cannot maintain per se, that we are measuring the same material; actually we are measuring samples of the material. We shall meet with this problem of sampling in later sections. When making replicate estimations on a patient's Red Cell Count, we are faced with problems of obtaining representative, homogeneous samples. It is

the central paradox of sampling, that we must ensure a 'representative' body of material from the population. But, if we do not 'know' what the population values should be (that is the reason for making the measurements), how can we be satisfied that we have achieved this aim? In this particular case we have 'sampled' from the patient - be that capillary or venous blood. In the laboratory we require to sample from this sample in order to make our determination(s). In most cases we make the assumption that the 'material' in each aliquot is the same. This leads us to believe that we are in fact making replicate measurements in the sense described above.

The measuring of the parameters of interest has, as we have indicated in the introduction and discuss in 3.5, two components - accuracy and precision. The precision of the measuring procedure chosen, will not usually be able to be altered, although it may vary due to reagents, temperature, humidity, technician fatigue etc. The precision of the estimate of the parameter of interest can be changed, however. By making replicate measurements on the specimen, the precision of the estimate will be increased.

It is assumed in most discussions that the measurements made are independent. This is a statistical use of the term, which can be defined in the following way. If we consider making a sequence of measurements $X_1, X_2, \ldots X_n$, we will have, after any one series of observations, particular values of the X_i, $i = 1,2,\ldots,n$ which we will refer to as x_i, $i = 1,2,\ldots, n$, i.e. the X_i are random variables and the x_i are the specific values of the random variables which have been observed in this case. If the conditions of the successive experiments (i.e. measurements), are strictly uniform, the probability of any specified event connected with say the i^{th}

experiment cannot be supposed to be in any way influenced by the results of the i-1 preceding experiments, then the measurements are called independent. For a rigorous treatment see e.g. Cramer (B 8) or Wilks (B 36).

It is reasonable to suppose that sequential measuring procedures, automatic or otherwise, will have a tendency to show dependence between successive measurements. But even in non-automated processes the readings are often influenced by previous readings.

Let us assume that a person makes five measurements on exactly the same material. The method allows measurement to the first decimal place with the next place estimated by the person measuring. The first four measurements made produced 15.15, 15.20, 15.10, and 15.10. If some dubiety is present as to whether the fifth measurement should be 15.20 or 15.25, most observers will probably record 15.20. This is due to the influence of the previous four measurements. In what follows, we will assume that the replicate measurements are independent, although the above warns us that this will not always be the case in practice.

As we have stated, the use of replicate measurements can lead to greater precision for our estimate. Consider the three following situations.

(i). Let us assume that n measurements are made, and further that they are independent and each has the same parent Gaussian distribution with mean μ and variance σ^2.

$X_i \sim N(\mu,\sigma^2) \quad i = 1, 2, \ldots, n$

The standard theory tells us that the arithmetic mean $\bar{x}$ has the following distribution

$\bar{x} \sim N(\mu,\frac{\sigma^2}{n})$

Further, it can be shown that $\bar{x}$ is the 'best' estimator of μ in this situation. It is the 'best' in the sense that it is unbiased, consistent, efficient and has minimum variance among all other unbiased estimators. If, for example, we were to choose to estimate μ by the median of the sample, then for large samples it can be shown that the median X_m has the distribution -

$X_m \sim N(\mu, \frac{1}{4n\phi^2(m)})$, where n is the sample size.

i.e. $E(X_m) = \mu$ and $V(X_m) \doteq \frac{\pi\sigma^2}{2n} > V(\bar{x})$

(ii). Let us now replace the assumption that the underlying distribution is Gaussian by some rather mild conditions on the random variable e.g. finite variance. Then the distribution of $\bar{x}$ is approximately Gaussian with mean μ and variance σ^2/n for large n. For reasonably 'well behaved' distributions, the approximation is good even for relatively small n. Again, $\bar{x}$ is the 'best' estimator of the population mean, μ.

The accuracy of the approximation to the Gaussian distribution is illustrated by the following example. The method to be described is used in some standard computer subroutines in generating random Gaussian deviates. A series of random numbers, each with a uniform distribution on the interval (0,1) can easily be generated. The mean of 12 such numbers is calculated. The distribution of such a quantity is such that the results obtained have, for many purposes, a sufficiently

good approximation to a Gaussian distribution.

(iii). Let us now consider a pathological example which shows that $\bar{x}$ need not necessarily be the 'best' estimator. If we define $Z = X/Y$, where X and Y are independent standard Gaussian random variables, then Z is a Cauchy random variable with a probability density function f(z), given by:-

$$f(z) = \frac{1}{\pi(1+z^2)} \qquad -\infty<z<\infty$$

It can be shown that if $Z_1, Z_2, \ldots, Z_n$ are independent Cauchy random variables with this density function, then $\bar{z}$ has the density f(z), i.e. the same density for the mean as any of the individual random variables. Further no moment exists - that is no finite mean nor finite variance. The 'best' estimator of the location parameter is not $\bar{z}$. The precision of this estimate does not increase as the sample size increases - it is not a consistent estimator. However, the median of the sample, Z_m, does estimate the location parameter and has asymptotic distribution given by

$$Z_m \sim N(0, \frac{\pi^2}{4n})$$

Fortunately the situation in (iii) is more of a theoretical problem than one which occurs commonly in practice. The remarkable result outlined in (ii) is the cornerstone of all practical statistics. It usually goes under the name of the Central Limit Theorem. It is not only a solace and comfort, but to some an insurance. However, we shall find that the results and logical extensions of the Central Limit Theorem must be seen in the light of the assumptions made,

no matter how general and mild they might look.

Consider a sequence of measurements made on a specific sample. We assume that μ_x is the 'true' but unknown value of the parameter which we require to measure. Let σ^2 be the precision of the method by which we measure the parameter. The sequence of random variables is $X_1, X_2, \ldots, X_n$, i.e. n replicate measurements. The usual assumptions are that the X_i's are independent and identically distributed random variables. We have discussed independence in the section above. The assumption of identical distributions is critical, especially in the case of routine laboratory methods. If we assume that each of the random variables has the Gaussian distribution $N(\mu_x, \sigma^2)$, this implies that each measurement x_i is an observation from a Gaussian population with mean μ_x and variance σ^2. However, in many situations, one or more of the x_i may look too large or too small to be an observation from such a population. In the experience of the analyst, such values are not 'likely' if the sequence of measurements were really replicate measurements. In this situation the belief in the assumption of identically distributed random variables is doubtful. The rogue values, called outliers, have been obtained by contamination or error, and can be considered not to be observations from the specified normal model. We have the following problems:-

(i). Can we identify which of the sequence of observations, if any, are outliers?

(ii). Can we produce a method of estimation which is not unduly affected by such observations when present?

These points, (i) and (ii), were made by Dixon (1953). We have above a model for the situation without outliers, but if we wish to model

the 'new' situation, we require to model the process which leads to the outliers. The model should allow for random outliers which occur infrequently. If the measuring process produced frequent outliers, it should not be a part of normal laboratory practice. Some suggestions as to the distribution that the outliers may come from are:-

$$N(\mu + d\sigma, \sigma^2) \ , \ N(\mu, \sigma^2(1+d'))$$

with d' and d, either constant or random. The major difficulty is of a practical nature. How can we determine a reasonable model for this contaminant distribution when the observations from it are mixed in with observations from another distribution, when the frequency of these observations is infrequent and perhaps most important, when the physical causes producing the outliers may be many and varied. This last point would suggest that there may in fact be more than one 'outlier distribution'. The suggested distributions above are only two suggestions from a large number of possibilities.

The theory of outlier detection has been developed in many sources. It is not proposed that a discussion of this theory be given here. However, certain points will be required in the following work. These will be discussed as we require them.

In section 2.3.1, we discuss briefly the effects that rounding of numerical results has on a clinical measurement. Measurement by matching is in section 2.3.2. Both situations occur frequently in the clinical laboratory. The discussion shows that numerical procedures, not only technical and medical procedures, affect the clinical results. This point is emphasised by the discussion of sections 2.3.3-5, which contains a resume of results in outlier theory. This

is yet another area in which the numerical treatment of clinical measurements can affect the values reported to the clinician.

2.3 Handling Numerical Results

2.3.1 The Effect of Rounding

In this section we discuss the effect of rounding on the measurements made. A clinician, technician or statistician must, of necessity, deal with rounded numbers - at least to the extent that the data are recorded values of continuous variables. (Of course, observations are also made on discrete variables, e.g. Red cell count, but even in these cases we can experience rounding. The more coarsely the data is rounded, the less accurately can it represent the 'observed facts'. Rules for rounding to a specific number of significant figures are well established in laboratories. The most common and most widely accepted are given by Eisenhart (1947) in the following form.

(i). If the first digit to be dropped is less than 5, the last digit retained shall be left unchanged.

(ii). If the first digit to be dropped is greater than 5, or is a 5 followed by digits greater than 0, the last digit retained shall be increased by 1.

(iii). If the last digit to be dropped is a 5 alone, or a 5 followed by 0's only, the last digit retained shall be rounded to the nearest even number.

Hildebrand (B 20) has made two important points. Firstly by rounding, we replace a number by an approximation with minimum error. Rounding is a different process from truncation, where we simply

discard unwanted digits. Secondly, the third rule will lead to a slight favouring of even numbers. With subsequent operations on these rounded numbers, we are less likely to require additional roundings. We shall adopt the rule given above in any rounding situation. We note that repeated use of this rule on the same number will not necessarily lead to the same final number as a single use of the rule will produce.

Very often in using statistics for analysis of data, we propose that the data can be assumed to be observations from a Gaussian distribution. The assumption regarding the measurements are that they are observations of a continuous random variable. But as we have suggested above, we have rounded the measurements, i.e. we have made the measurements discrete. Let us consider the effects of such a rounding procedure.

Consider measuring the haemoglobin level of 'normal' males, aged 20-30 years. The distribution of values can be considered to be Gaussian, with mean 16 g/dl and standard deviation 2.0 g/dl. On a machine such as the Coulter Counter Model S (for a description see section 3.4), we record measurements to 1 decimal place g/dl. Technically, we say that the rounding lattice is,

(... 13.7, 13.8, 13.9, 14.0, 14.1, 14.2, 14.3,....)

We formulate the basic problem in the following way. The continuous random variable X, has a Gaussian distribution with mean μ and variance σ^2. The random variable is rounded to the nearest value X_R. The rounding lattice is,

$$(\ldots\ldots, x_{Ri-1}, x_{Ri}, x_{Ri+1}, \ldots\ldots)$$

That is, the possible values of the random variable X_R are x_{Rj}, $j = 0, \pm 1, \pm 2, \ldots$. We have in fact made discrete

the random variable X. The lattice is centred on the x_{Ri}, with interval of width $2a\sigma$. The lattice is positioned in such a way that the interval containing μ is centred at $\mu + D\sigma$, $-a<D\leq a$. We can assume, without loss of generality, that $\mu = 0$ and that $\sigma^2 = 1$. The distribution of X_R is given by:-

$$P(X_R = x_{Ri}) = P(x_{Ri} - a \leq X \leq x_{Ri} + a)$$

$$= \int_{a_{Ri}-a}^{x_{Ri}+a} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

where $x_{Ri} = D + 2ia \quad i = 0, \pm 1, \pm 2, \ldots\ldots$

Given D and a, the distribution of X_R is completely specified. The problem is, that in practice, although a will be known, D will never be known.

If we consider such a situation and suppose that we calibrate the measuring process such that the description above is apt, then we note the following points:-

(i). The random variable X_R is discrete.

(ii). The expected value of X_R is not necessarily 0, (i.e. not necessarily the 'true' population value of the mean of X).

(iii). The variance of X_R is always greater than the variance of X.

In most practical situations, the range of X_R and X will not be infinite. Often we truncate the ranges to, say, $(\mu-\delta\sigma, \mu+\delta\sigma)$ for some positive value of δ usually greater than 3.5. The distribution of X_R has been given, for some specific values of D and a, by Speckman (1964).

In the above, we have assumed that the rounding lattice is fixed at a certain specific position with respect to the true mean of X. As we have indicated, in practice we can never determine the value of D which describes the positioning. We have also assumed that the positioning is fixed for all measurements made with the process. In practice, due to recalibrations etc., the value of D will be a random variable. Let us assume that the distribution of D is symmetric on the interval (-a, a). We denote by $\bar{x}$ the mean of a sample taken on the random variable X, and by $\bar{x}_R$ the sample mean of observations on the random variable X_R. Then we note the following point.

(i). Both $\bar{x}$ and $\bar{x}_R$ are unbiased estimators of the population mean of X.

(ii). The variance of $\bar{x}_R$ is $(1 + a^2/_{12})$ times the variance of $\bar{x}$.

(iii). The skewness of the distribution of $\bar{x}_R$ never exceeds that of $\bar{x}$.

(iv). The kurtosis of the distribution of $\bar{x}_R$ is always less than that of $\bar{x}$.

We note that in this situation, $\bar{x}_R$ is an unbiased estimate of μ, unlike the one-off case discussed above. Further, we note that the variance of the measured population is affected by the rounding of the measurements. The effects of the increase in variance, may be important, e.g. when considering the construction of 'normal ranges', see section 5.2.

Let us define the random variable X as the haemoglobin concentration, g/dl, in a specific male's (healthy) blood specimen. Then X is a continuous random variable. Let us suppose that a

Coulter-S counter is employed in the determinations. We assume that the device is 'in control'. The Coulter-S results are replicate observations on a discrete random variable, X_R. The random variable X_R is the result of rounding the variate X.

Using the same notation as in the above, the rounding lattice has width 2a = 0.1, since the measurement is reported to the nearest tenth g/dl. From the Coulter-S 4C assay sheet, the standard deviation for repeat analysis is 0.1 g/dl. Let $\bar{x}_R$ be the mean of the n repeat measurements on X_R and let $\bar{x}$ be the mean of the corresponding n measurements on the continuous random variable X. If we assume that X has a Gaussian distribution with mean 16.0 g/dl, then let

$$\bar{x} \sim N(16.0, \frac{\sigma^2}{n})$$

$$E[\bar{x}_R] = 16.0 \text{ and } V[\bar{x}_R] = (1 + \frac{0.05^2}{3}) \frac{\sigma^2}{n} = 0.1^2 \text{ (with } n = 1)$$

That is $\sigma^2 = (0.09996)^2$

Thus, in this particular case, the effect of the rounding is negligible in terms of its effect on the distribution of results in the population. In fact this is true for all the reported values measured by the Coulter Counter models.

Although we have shown in this situation, that the rounding experienced on the Coulter Counter model S is negligible with respect to distributional effects, this may not be true for some other facet. For example, the effects of rounding may not be negligible in certain quality control procedures, see chapters 5 and 6. Rounding may also effect procedures for the detection of outliers, see section 2.3.3.

Finally we note that the above situation is the simplest which we could encounter. It will often be the case that we wish to measure functions of discretised variables, e.g. $Z_1 = X/Y$ or $Z_2 = XY$ where

either X or Y or both may have been rounded as described above. The behaviour of Z_1 and Z_2 is not easy to determine analytically. If for example we assume that both X and Y have Gaussian distributions the resulting distribution for Z_1 is related to the Cauchy distribution (see 2.2.2). Approximations for the mean and variance of functions of random variables are given in Ku (B 24) and will not be considered further here.

2.3.2 Measuring by Matching

In the section above, we have considered a continuous random variable which is made discrete by the measuring process. In such a case the process offers us those values of the rounding lattice as the only answers to our measuring. Thus, if we are asked to measure a specimen, we in reality judge which of the rounding points is closest to the true value of the specimen. This situation is frequently found in qualitative measurements e.g. colour matching.

Let us once again consider discussing this problem in general terms first, and then give an indication of how it may affect us in practice. We define terms of interest in the same manner as Youden, Connor and Severo (1959). The lattice points will be called the standard values. The value which is assigned to the specimen being measured is the closest standard value to the measured value. This will be called the assigned standard value. We note that this is not necessarily the standard value closest to the 'true' value. For simplicity we take the true, but unknown value to be 0. (This does not affect the generality of the results). Let the measured value be Y, and the assigned standard value be X. The quantity X-Y is the difference between the measured quantity and the assigned value, which we represent by Z. Then

$$X = Y + Z$$

The lattice of standard values is fixed and the intervals are of width 2a. One of two assumptions can be made:-

(i). The true value is uniformly distributed in the interval (-a,a).

(ii). The true value is fixed at 0 and the grid is imposed at random.

Eisenhart (1947) has shown that these assumptions are in fact equivalent. We shall use the formulation of (ii).

We are primarily interested in two probabilites which Youden et al put in the following forms.

(i). The probability that an unknown specimen will be assigned the standard value which is closest to its true value.

(ii). The probability that two independent measurements on the same specimen will result in the same assigned value, regardless of whether it is the closest standard value or not.

We assume that the measurement process is 'in control' and denote the probabilites in (i) and (ii) by P(A) and P(B) respectively. We assume that Y has a Gaussian distribution with mean 0 and variance σ^2 and that Z has a uniform distribution on (-a,a). As Youden et al show, the probabilities P(A) and P(B) are given by the following.

$$P(A) = \left[\Phi\left(\frac{2a}{\sigma}\right) - \Phi\left(-\frac{2a}{\sigma}\right)\right] + \frac{\sigma}{a}\left[\Phi\left(\frac{2a}{\sigma}\right) - \phi(0)\right]$$

$$\text{and } P(B) = \left[\Phi\left(\frac{\sqrt{2}a}{\sigma}\right) - \Phi\left(-\frac{\sqrt{2}a}{\sigma}\right)\right] + \frac{\sqrt{2}\sigma}{a}\left[\phi\left(\frac{\sqrt{2}a}{\sigma}\right) - \phi(0)\right]$$

$$\text{where } \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}x^2\right) \text{ and } \Phi(x) = \int_{-\infty}^{x} \phi(t)dt.$$

Consider, for example, measuring the haemoglobin level of a patient specimen using the Coulter-S. As we have indicated in the last section, the value will be reported to the nearest 0.1 g/dl. The standard deviation σ, in the expression above, is the process precision. Then a = 0.05 and we have:

σ = 0.05	P(A) = 0.6151	P(B) = 0.4890
σ = 0.10	P(A) = 0.3688	P(B) = 0.3566
σ = 0.20	P(B) = 0.1958	P(B) = 0.1101

As we expected, the probabilities depend on the ratio of the width of the rounding lattice to the process precision. The stated precision for the Hb channel (by Coulter) is 0.1 g/dl. Thus, just slightly more than a third of the time, when the device is in control, do we have reported the nearest standard value. Further, with approximately the same frequency we have replicate measurements reported to have the same assigned values.

The situation regarding P(B) is, in practice, much more complicated than as indicated by the analysis. The assumption regarding the process as in control is paramount to the analysis. If control (calibration etc.) has been lost between replicate readings the estimate, P(B), is almost certainly an upper bound to the actual probability.

The discussion of the problem of measuring by matching, should highlight the effect of 'rounding' i.e. the problem of a continuous random variable being made discrete by the measurement process. Even with none of the practical problems admitted e.g. inhomogeneity of samples, carryover, drifting, etc., the analysis shows that we can expect in a small number of times the nearest standard value to the true value to be reported, and with approximately the same frequency,

replicates to have the same reported values.

We have discussed two particular situations - the rounding of numerical measurements (either by a machine or by laboratory personnel) and the measuring of a parameter by matching - which occur in clinical laboratories. The discussion should show that even accepted and commonly used laboratory procedures which seem intuitive and at first sight reasonable can affect the quality of the laboratory's work. The extent of such effects can be gauged. In some particular situations these effects could be substantial. The examples which have been used in the discussion are related to the Coulter Counter model S.

Having shown that some procedures in the handling of data could cause a decreased quality in work, we now investigate the effects of outliers, i.e. spurious or rogue results on the quality of clinical results.

2.3.3 Outliers

In this section we discuss briefly how outliers in data may be detected, and what may be done upon their detection or what may be done practically in their presence. Before this, we consider how outliers might arise in a haematology laboratory.

The basic problem requires us to identify outliers in a sample of size n. The observations comprising this sample should, by assumption, have come from the same parent distribution. The smallest sample which can be investigated for outliers with no concomitant information is a sample of size three. As we point out later this basic situation, three replicate measurements, occurs frequently in the laboratory.

Let us consider that three measurements x_1, x_2, and x_3 have been made. These measurements are replicate measurements on the

same material. We order these in ascending order of magnitude which gives z_1, z_2, z_3. The range of the sample is w, which is the difference between the largest and the smallest observation, i.e.

$$w = z_3 - z_1$$

Denote by x' and x'', (x'' < x') the closest observations in the sample. Further, let the ratio of the difference between the two closest observations and the sample range be y.

$$y = \frac{x' - x''}{z_3 - z_1} \geqslant 0$$

Assume that the underlying parent distribution is Gaussian. We denote by F(z) and f(z) the distribution function and the density function of y. It can be shown that these are:-

$$f(z) = \frac{3\sqrt{3}}{\pi} \frac{1}{1-y+y^2} \quad 0 \leqslant y < \tfrac{1}{2}$$

$$F(y) = \frac{6}{\pi} \tan^{-1} \left(\frac{2y-1}{\sqrt{3}}\right) + 1$$

Hence we can construct the probabilities given in table 2.1

TABLE 2.1

k	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{11}$	$\frac{1}{16}$	$\frac{1}{21}$
P(y⩽k)	0.6369	0.4634	0.3631	0.2983	0.1572	0.1066	0.0806

As Lieblein (1952) has indicated, this ratio is not a good criterion to use for the rejection of values as outliers. Even when the parent distribution is Gaussian, we can expect the outermost observation to be five times as distant from the middle one as the remaining one in about 30% of cases. This occurs even when there are no outliers present. The situation is exacerbated when the observations are rounded, due to the possibly high probability of two values being reported as equal, see section 2.3.2.

Lieblein has also shown that the behaviour of the statistics, y, is almost the same if the underlying parent distribution is rectangular or right triangular. In fact this work should indicate to us that such a criterion is not good for determining the presence of outliers.

If this procedure is not used, what other criterion may be used? Let us assume that the parent distribution is Gaussian, with known variance σ^2. Consider the quantity, w/σ, which is the 'studentised range'. Pearson and Hartley (B31) have constructed tables for this statistic. For a sample of size three, the mean value of w/σ is 1.693, and the standard deviation is 0.888. A rule, sometimes proposed for the rejection of outliers is to reject any observation which causes the range of the sample to exceed three times the standard deviation σ. It can be seen from these tables that this will happen in about 9% of cases, even when no outliers are present. (Note we assume that the variance σ^2 is known. If this requires to be estimated, obviously the precision of statements is lessened).

The above should have the effect of making us less confident about rejection of observations as outliers. But reject we must in cases where gross errors appear to have occurred. The point is: when can we be sure that an observation is an outlier? The answer is, of course, never - unless some other pertinent information is present. The literature concerning outliers investigates procedures for increasing the probability of correctly dealing with suspected observations.

Procedures for the detection and subsequent rejection of outliers are numerous. Chauvenet in the middle 19th century, proposed a scheme which is used in a slightly modified form today. (We shall see it used in the DHSS (LDAG) and BCSH Haematology Quality Control Trials -

see 6.8). Many other procedures have been proposed. For a discussion of some of these, see Guttman and Smith (1969), Grubbs (1950)(1969), Anscombe (1960), Dixon (1953), Dixon (1962) and Anscombe and Barron (1966).

Basically, as we have indicated, two situations can occur. First, we may have 'obvious' errors i.e. gross errors. Often these can be attributed to transcription errors e.g. reporting 41 g/dl in place of 14 g/dl for haemoglobin concentration. A gross error such as this can be rejected out of hand - for clinical reasons.

Secondly, the errors may not be gross and may not even be such that they are even 'suspicious'. In this case the outliers, if indeed that is what they are, cannot be rejected out of hand. These are the situations where outlier rejection procedures are particularly required.

Although detection and subsequent rejection of outliers is important in the clinical laboratory, we propose that in place of developing such rejection procedures, we attempt to construct procedures for handling experimental measurements, which are robust in the presence of outliers. The clinical measurements are estimates of the true clinical value of the parameter being measured. In the following we discuss procedures for handling replicate measurements which yield estimated values which are not too seriously affected by outliers if present. The degree to which the estimated values are affected by outliers obviously depends on the actual procedures employed. In the following we discuss some such procedures.

2.3.4 Estimation of Clinical Parameters in the Presence of Outliers.

Although we shall consider in this section outliers with respect to replicate measurements, other situations in the clinical

laboratory, can produce 'outliers'. This topic is discussed further in section 6.4. For the present, our description will be in terms of replicate measurements made on the same clinical material. Obviously, outliers such as the gross mistakes described at the end of the last section, should be detected in all clinical situations. However, more sophisticated procedures to detect or provide robust estimation may be required to deal with rogue observations which cannot be rejected out of hand. The types of situation in which outlier detection may be advantageous are accurate methods and/or methods which have complex sub-procedures, e.g. such as microbiological assays.

In the following we shall discuss two models which have been used to simulate outlier situations. Some commonly used procedures for estimation in the presence of outliers are discussed and their behaviour indicated for the models which we describe. Finally we comment on a procedure of triplicate counting which is found in the Coulter-S counter (to be described in section 3.4).

Consider replicate measurements made on the same sample, i.e. the same clinical material. (We shall assume that even in the case where the initial patient specimen has been split into smaller volumes, that we have the 'same sample'). The measurements made on these different smaller volumes with respect to the same clinical parameter are, for the present, considered to come from the same parent Gaussian distribution. Let the unknown true value of the i^{th} patient specimen be μ_i. The j^{th} replicate on the i^{th} patient specimen is denoted x_{ij}. For each measurement x_{ij} there is a corresponding random error ε_{ij}, which is assumed to have a Gaussian distribution with zero mean and variance σ^2. We write:-

$$x_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, 2, \ldots, n, \quad j = 1,2,3, \ldots r.$$

In this case $\bar{x}_i$ is the 'best' estimator of μ_i, (see section 2.2.2).

The model covers cases in which either the same instrument is used to make each of the replicate measurements (and is in control), or for example, where a number of test-tubes are used under the same experimental conditions - say in the estimation of serum vitamin B_{12} or of serum folate by microbiological methods.

The model described above infers that any differences between replicate measurements is solely caused by measurement error. We now make allowance for an infrequent disturbance which produces outliers in the replicate measurements. We propose the following model:-

$$x_{ij} = \mu_i + \beta_{ij} + \epsilon_{ij} \qquad \begin{matrix} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, r \end{matrix}$$

The same notation is used as before, with the same distributional assumptions. Two different sets of assumptions can be made concerning the term β_{ij}.

(a). β_{ij} can be expressed as $b\delta_{ij}$ where b is a constant and δ_{ij} is such that $P(\delta_{ij} = 1) = d$ and $P(\delta_{ij} = 0) = 1 - d$. We shall call b the bias and 100d the percentage contamination rate.

(b). β_{ij} is a random variable such that there is a probability p that the bias is zero. The random variable can take on other values, positive or negative, with probability 1-p.

We shall discuss each of these situations in turn.

(a). The situation described by the model given in (a) will be called the constant bias model. The contribution of the bias is constant when present. The percentage contamination rate describes the frequency with which we can expect contamination. In practice, we expect the contamination rate to be low; certainly small enough to give a small probability that, for example, no more than 1 outlier is contained in a set of three replicates. (We note that if the defective rate is d the probability that more than 1 outlier is present in a set of three replicates is $3(1-d)d^2 + d^3$. Thus for example when d = 0.2 and d = 0.1 the probability of more than one outlier are 0.104 and 0.028 respectively).

We note that the expected value and variance of the mean of r replicates, $\bar{x}_{i.} = \frac{1}{r} \Sigma x_{ij}$, is

$$E(\bar{x}_{i.}) = \mu_i + bd \text{ and } V(\bar{x}_{i.}) = \frac{\sigma^2}{r} + b^2 \frac{d(1-d)}{r}$$

In our illustrations we will consider b > 0. It will be clear that the replicate mean is a biased estimator of the true specimen value (on average). As we have indicated, this motivates us to consider procedures for dealing with the outliers.

(b). We shall call the model described by (b) the variable bias model. It is unlikely in practice that the factor(s) causing the outlier effect is constant. For example, the effect of detergent on the estimation of serum folate or serum vitamin B_{12} by microbiological methods is dependent on many things, not least of all on the amount of detergent present. In the model described, we make allowance for an infrequent random bias.

The random variable β_{ij}, has a mixed form, which as we have already indicated prescribes a probability p to the event that the random variable is zero. (We note that 100 (1-p) is analogous to the contamination rate of (a).

In practice it is difficult to know what type of distribution the random variable β_{ij} has. Let us assume, by way of an example, that when bias is present (i.e. outlier bias) it is such that $\beta_{ij} > 0$. Furthermore, let us assume that β_{ij} has a Gamma form, so that

$$P(\beta_{ij} < 0) = 0$$

$$P(\beta_{ij} = 0) = p$$

$$P(\beta_{ij} > b) = p + (1-p) \int_0^b (p\lambda)^n \frac{x^{n-1} 1^{-\lambda p x}}{\Gamma(n)} \, dx$$

The mean and variance of r replicate measurements is

$$E(\bar{x}_i) = \mu_i + \frac{n(1-p)}{p\lambda} \quad \text{and} \quad V(\bar{x}_i) = \frac{\sigma^2}{r} + \frac{n(1-p)}{r(p\lambda)^2}$$

where μ_i is the true value of the specimen replicated and σ is the precision of the measuring process.

Even in this rather simple case we have many more possibilities to consider as compared with the constant bias model. Later in this chapter, section 2.3.5, we consider the behaviour of some commonly accepted estimating procedures which have been proposed for use in the presence of outliers. We shall consider only model (a). It is thought that this will give sufficient indication of the behaviour of these procedures, without entering into hypothetical arguments concerning the distributional nature of the random bias.

As we have stated previously, we are interested in firstly deciding whether outliers are present in replicate measurements,

and secondly in constructing robust procedures for estimating the true clinical values of the patient specimens, given that outliers may be present.

2.3.5 Common procedures for Estimation in the Presence of Outliers

We consider some suggested and accepted procedures for estimation of the true value of the clinical parameter in the presence of outliers in replicates. We specifically discuss triple replicates (although we compare the results with a double-replicate situation). Each estimator will be numbered, so that they may be easily identified in the section which deals with the results of some computer simulations. Throughout this section we will assume that the precision of the measuring process is known, and is σ. In practice, this is not the case and we require to replace σ by s the estimated precision.

(i). Two replicate measurements, x_1 and x_2, are made. The estimate is the mean of the measurements

i.e. $y_1 = \dfrac{x_1 + x_2}{2}$

(ii). Two measurements, x_1 and x_2, are made. If their difference is less than a predetermined quantity, d, the estimate is the mean of x_1 and x_2. If the difference is greater than d, another replicate measurement, x_3, is made and the estimate is the mean of the closest two of the three measurements. (d is usually expressed in terms of σ, and frequently is in the range $(\sigma\ , 3\sigma)$.

$$\text{i.e. } y_2 = \begin{cases} \dfrac{x_1+x_2}{2} \text{ if } |x_1-x_2| \leqslant d \\ \dfrac{x_1+x_3}{2} \text{ if } |x_1-x_2| > d \text{ and } |x_1-x_3| \leqslant |x_2-x_3| \\ \dfrac{x_2+x_3}{2} \text{ if } |x_1-x_2| > d \text{ and } |x_1-x_3| > |x_2-x_3| \end{cases}$$

(iii). Three replicate measurements, x_1, x_2 and x_3 are made. The estimate is the mean of the three measurements

i.e. $y_3 = \dfrac{x_1 + x_2 + x_3}{3}$

(iv). Three replicate measurements, x_1, x_2 and x_3 are made. The estimate is the median.

The measurements x_1, x_2 and x_3 are ordered in ascending order of magnitude to give z_1, z_2, z_3 (where $z_1 \leqslant z_2 \leqslant z_3$). The median is z_2.

i.e. $y_4 = z_2$.

(v). Three replicate measurements x_1, x_2 and x_3 are made. The estimate is the mean of the two closest measurement. We denote by x' and x'' the two closest measurements. ($x' \geqslant x''$)

i.e. $y_5 = \dfrac{x' + x''}{2}$

(vi). Three replicate measurements x_1, x_2 and x_3 are made. If the range of the replicates (i.e. largest minus the smallest, $z_3 - z_1$ in notation of (iv)) is less than d, frequently 3σ, then the estimate is the mean of the three observations. If the range is greater than d, the estimate is the mean of the two closest observations.

i.e. $y_6 = \begin{cases} \dfrac{x_1 + x_2 + y_3}{3} & \text{if } z_3 - z_1 \leqslant d \\ \dfrac{x' + x''}{2} & \text{if } z_3 - z_1 > d. \end{cases}$

(vii). Three replicate measurements x_1, x_2 and x_3 are made. The mean of the measurements, $\bar{x}$, is calculated. If the difference between the smallest and $\bar{x}$, and the largest and $\bar{x}$ are both less than d, the estimate is the

mean $\bar{x}$. (Once again d is frequently taken as 3σ). If these differences are both greater than d, the estimate is the sample median. If either difference is greater than d while the other is not, the mean of a modified set of measurements is taken. (The modification essentially makes less extreme the 'outlier'.)

$$y_7 = \begin{cases} \bar{x} & \text{if } (\bar{x}-z_1),\ (z_3-\bar{x}) < d \\ z_2 & \text{if } (z_2-z_1),\ (z_3-z_2) > d \\ \dfrac{z_1+2z_2+k}{3} & \text{if } z_3-z_2 > d > z_2-z_1 \\ \dfrac{-k+2z_2+z_3}{3} & \text{if } z_3-z_2 < d < z_2-z_1 \end{cases}$$

Anscombe and Barron (1966) have discussed which value of k should be used to give the 'best' procedure. We shall consider values in the range $(\sigma, 3\sigma)$.

(viii). Two sets of three replicate measurements x_{11}, x_{12} and x_{13}, and x_{21}, x_{22} and x_{23} are made. The mean of the two closest observations in each set is calculated. If these means differ by less than d, the mean of these two means is the required estimate. If the means differ by more than d, which has a value similar to (vi), a further set of three replicate measurements x_{31}, x_{32} and x_{33} is made. The mean of the closest two observations in this set is calculated. The estimate is taken as the mean of the two closest means.

$$y_{81} = \frac{x_1' + x_1''}{2}, \quad y_{82} = \frac{x_2' + x_2''}{2} \text{ and if necessary } y_{83} = \frac{x_3' + x_3''}{2}$$

$$y_8 = \begin{cases} \frac{y_{81}+y_{82}}{2} \text{ if } |y_{81}-y_{82}| \leq d \\ \frac{y_{81}+y_{83}}{2} \text{ if } |y_{81}-y_{82}| > d \text{ and } |y_{81}-y_{83}| \leq |y_{82}-y_{83}| \\ \frac{y_{82}-y_{83}}{2} \text{ if } |y_{81}-y_{82}| > d \text{ and } |y_{81}-y_{83}| > |y_{82}-y_{83}| \end{cases}$$

Now we consider the data handling procedures (i)-(viii) described above in conjunction with the constant bias model (a) described in section 2.3.4.

2.3.6 Results of Simulation of a Measuring Process and Estimating Procedures

In this section, we present the results obtained by simulating 'triplicate' measurements and employing each of the data handling procedures (i)-(viii) described in 2.3.5. The measurements were standard Gaussian deviates (generated by the Nag routine G05ADF). Three parameters are varied; the percentage contamination rate, the constant bias and the filter size (this is a parameter required of some of the procedures). A flowchart of the computer program is given in figure 2.1, and a listing of the program is given in Appendix 1. The results are presented at the end of this chapter in tables. The tables show the mean and standard deviation of 100 sets of triplicate measurements. Furthermore, the quantity 'toterr' represents the total error. This is a combination of the bias and lack of precision of the procedure, and in this case is $\sqrt{\frac{1}{n-1}\Sigma x_{ij}^2}$, where x_{ij} is the estimated true value of the i^{th} 'specimen' using the j^{th} procedure.

We comment briefly on the expected benefits and the practicality of using these different estimators. The comments made come from two sources - simulation experiments, which have been described above, and published work by a number of authors. The simulation

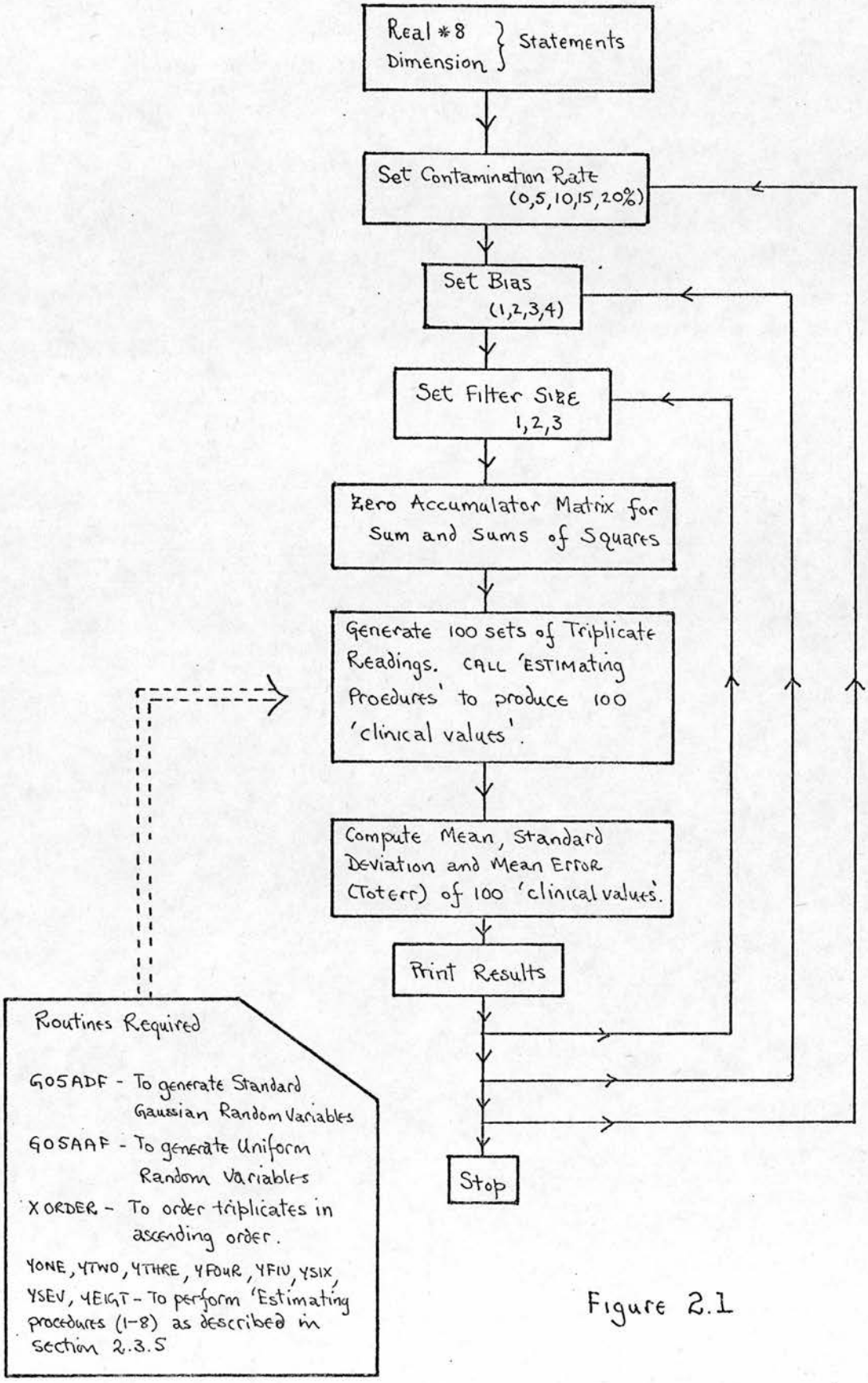

Real *8
Dimension
Statements
Set Contamination Rate
(0,5,10,15,20%)
Set Bias
(1,2,3,4)
Set Filter Size
1,2,3
Zero Accumulator Matrix for
Sum and Sums of Squares
Generate 100 sets of Triplicate
Readings. CALL 'ESTIMATING
PROCEDURES' to produce 100
'clinical values'.
Compute Mean, Standard
Deviation and Mean Error
(TOTERR) of 100 'clinical values'.
Print Results
Stop
Routines Required
G05ADF - To generate Standard
Gaussian Random Variables
G05AAF - To generate Uniform
Random Variables
XORDER - To order triplicates in
ascending order.
YONE, YTWO, YTHRE, YFOUR, YFIV, YSIX,
YSEV, YEIGT - To perform 'Estimating
procedures (1-8) as described in
section 2.3.5

Figure 2.1

was carried out to compare simultaneously the capabilities of the different estimators. Published work by Anscombe and Barron (1966), Wilke (1966), Lieblein (1952) and Veale and Huntsberger (1969), deal with specific sets of the estimators which we have labelled y_1, y_2, ..., y_7.

Before we turn to the discussion, it is pertinent to quote Anscombe and Barron on the subject of determining, using simulation experiments, which of a set of estimators is the 'best'.

'No set of calculations such as ours can settle the matter beyond dispute, because we do not know that reality is well represented by the distribution assumptions that we have used, nor by any other distribution assumptions that we might have used instead.'
This is reminiscent of our previous note, that the specific clinical situation under investigation deserved particular examination.

When no contamination is present, basic statistical theory states that the mean of m observations is 'better' than the mean of n observations, (m>n). Further, if m = 3 and n = 2, both give better estimates, on average, than the median of a sample of three. Hence when no contamination is present, y_3 should be preferred to either of y_1 or y_4. In a comparison between y_5 and y_2, Lieblein has shown that y_5 has a larger standard error than y_2. When no contamination is present, y_7 has a standard error only slightly larger than y_3.

We note from the results of the computer simulation, see table 2.2, that for large biases, and reasonable filter size i.e. greater than 1 standard deviation, y_8 performs on average quite markedly better than any of the other estimators. Further, on average, y_8 performs better than the other estimators in general. This is not a surprising result- y_8 can use up to three times as much information as the others.

As the bias increases, and the frequency of contamination reaches 20%, the 'mean of two and three' i.e. y_1 and y_3 fare badly in comparison with the 'mean of the two closest' i.e. y_5. This should be expected, since the contaminant, usually a maximum of one in the sample of three, is discarded, and the remaining two averaged.

Anscombe and Barron conclude that y_7 is usually better than y_6. As we see from table 2.2, if we compare 'toterr' this conclusion is supported.

For estimators, y_2, y_6, y_7, y_8, relying on the quantity d, the filter size, we must have d large enough to prevent discarding 'good' observations, but small enough to allow for the discarding of outliers. In most practical situations d should be between two to three times the process precision, and most likely to be approximately three times. This is borne out by examination of the behaviour of the estimators with respect to variation in filter size.

In general, the conclusions which can be made are that y_8, y_7 and y_6 usually fare well, especially with large contamination. Of these estimators, as we have indicated above, y_8 is 'best'. The estimators y_5, y_4, y_3 and y_2 have varying degrees of success depending on the contamination. It is interesting to note from table 2.2, that y_3, the mean of three observations does not fare as badly as we might expect. Obviously this is peculiar to the parameter values which have been used, i.e. 20% contamination and up to 4 times the process precision as bias.

In the clinical laboratory, any comparison of such estimators must take into account the computing ability of staff and laboratory, and the ease whereby arithmetical errors can occur in the computations. Further, any estimator used is obviously dependent on the accuracy and

and precision of the measuring process. The accuracy and precision demanded of the process is dependent on the clinical situation. If small discrepancies in estimated value and true value can be tolerated, i.e. are not clinical significant, then we do not require the sophisticated procedures e.g. y_8. Any estimators used must be tailored not only to the clinical situation, but also the actual measuring method. The estimating procedures should be understandable to the operators and unless mechanised, should contain as few calculations as possible. Further, if the estimator is to be used with any justification, it must do away with value judgements made by the operator(s), which is one of the main reasons for its use.

Before leaving this topic of rejection of outliers, we discuss a situation encountered on the Coulter-S counter. (For a description of this device see section 3.4). As we shall see the clinical parameters which the device reports as 'primary measures' i.e. WBC, RBC, Hb and MCV are based on 'triplicate determinations.' The aspirated blood is split in two parts and one part is used to determine WBC and Hb the other being used to determine RBC and MCV. Each of the clinical values reported is based on a triple count. The basic situation for each of these determinations is essentially different from the 'replicate measurement' just described.

Let us consider a single clinical parameter. This parameter is determined by three different counters, each of which can be independently calibrated. Let y_{ij} denote the j^{th} replicate on the i^{th} specimen. Then we suggest that we can write:-

$$y_{ij} = \alpha_j + \beta_j \mu_i + \delta_{ij} + \varepsilon_{ij} \qquad \begin{matrix} i = 1, 2, \ldots. \\ j = 1, 2, 3. \end{matrix}$$



In this case, α_j and β_j are parameters describing the calibration of the j^{th} counter. The random errors are denoted ε_{ij}. The true value of the i^{th} specimen is μ_i. The bias, not due to calibration is δ_{ij}. We note that y_{ij}, $j = 1, 2, 3$ are not replicates in the true sense (not unless ($\alpha_j = \alpha$, $\beta_j = \beta, j = 1, 2, 3$). To detect non-zero biases i.e. 'outliers', we must rely on similar methods are proposed above. However, the previous methods were based on the assumption that the only differences, other than random error between 'replicates' was caused by the bias and not calibration differences. Thus, although we could employ the previous methods, the properties of the methods will be different.

In fact, automatic outlier detection methods have been built-in to the Coulter-S. For WBC, RBC and Hb a count is rejected if it differs by more than 3 standard deviations (channel precision) from the others. For MCV the rejection criterion is based on a percentage of the mean value. (We note that the Coulter-S manuals are not specific on these points).

TABLE 2.2 (PAGE 54 - PAGE 62)

THE FOLLOWING IS A PHOTOCOPY (REDUCED)
OF THE OUTPUT OBTAINED FROM THE
PROGRAM LISTED IN APPENDIX 1.
FOR AN EXPLANATION SEE SECTION 2.3.6.

CONTAMINATION RATE - 0 PERCENT

BIAS - 1 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.010	-0.001	0.008	-0.034	-0.081	-0.077	-0.018	-0.111
ST.DEV	0.729	0.809	0.656	0.741	0.860	0.845	0.700	0.588
TOTERR	0.729	0.810	0.657	0.742	0.864	0.848	0.700	0.599

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.037	0.050	0.012	-0.008	-0.024	0.006	0.011	0.063
ST.DEV	0.674	0.749	0.567	0.628	0.704	0.669	0.568	0.536
TOTERR	0.675	0.751	0.567	0.628	0.704	0.669	0.568	0.540

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.050	-0.044	-0.039	0.002	0.039	0.033	-0.038	0.017
ST.DEV	0.726	0.734	0.488	0.532	0.646	0.560	0.488	0.578
TOTERR	0.727	0.735	0.490	0.532	0.647	0.561	0.490	0.578

BIAS - 2 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.016	0.001	0.026	0.036	0.022	0.024	0.026	0.080
ST.DEV	0.691	0.765	0.595	0.681	0.795	0.782	0.615	0.573
TOTERR	0.691	0.767	0.595	0.682	0.795	0.783	0.616	0.579

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.081	-0.077	-0.087	-0.134	-0.151	-0.164	-0.094	-0.097
ST.DEV	0.730	0.797	0.604	0.712	0.840	0.780	0.606	0.616
TOTERR	0.734	0.741	0.611	0.724	0.853	0.796	0.614	0.624

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.119	-0.114	-0.101	-0.051	-0.026	-0.107	-0.101	0.043
ST.DEV	0.809	0.802	0.600	0.746	0.905	0.699	0.600	0.673
TOTERR	0.818	0.810	0.608	0.748	0.905	0.707	0.608	0.675

BIAS - 3 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.023	-0.025	-0.047	0.015	0.033	0.041	-0.018	0.030
ST.DEV	0.703	0.784	0.577	0.721	0.870	0.857	0.647	0.607
TOTERR	0.704	0.784	0.579	0.721	0.871	0.858	0.647	0.608

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.019	-0.050	0.018	0.001	0.051	0.037	0.021	0.011
ST.DEV	0.661	0.671	0.574	0.642	0.768	0.697	0.576	0.532
TOTERR	0.661	0.673	0.575	0.642	0.770	0.698	0.576	0.532

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.039	-0.078	-0.050	-0.092	-0.079	-0.064	-0.050	-0.055
ST.DEV	0.652	0.675	0.570	0.697	0.845	0.652	0.570	0.581
TOTERR	0.655	0.680	0.572	0.703	0.848	0.655	0.572	0.586

BIAS - 4 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.093	-0.084	-0.051	-0.036	0.013	0.016	-0.027	0.007
ST.DEV	0.732	0.730	0.623	0.659	0.765	0.747	0.612	0.551
TOTERR	0.738	0.735	0.625	0.660	0.765	0.748	0.613	0.551

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.087	-0.071	-0.028	0.022	0.048	0.026	-0.022	0.013
ST.DEV	0.623	0.638	0.536	0.610	0.724	0.665	0.538	0.569
TOTERR	0.629	0.642	0.536	0.610	0.725	0.666	0.538	0.569

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.016	0.003	-0.043	-0.092	-0.122	-0.049	-0.043	-0.058
ST.DEV	0.678	0.691	0.548	0.660	0.810	0.600	0.548	0.532
TOTERR	0.678	0.691	0.550	0.666	0.819	0.602	0.550	0.535

CONTAMINATION RATE - 5 PERCENT

BIAS - 1 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.012	-0.044	0.041	0.001	-0.053	-0.055	0.017	0.111
ST.DEV	0.800	0.836	0.668	0.713	0.839	0.837	0.666	0.637
TOTERR	0.800	0.837	0.669	0.713	0.841	0.838	0.666	0.647

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.108	0.066	0.045	0.009	-0.029	-0.004	0.027	0.059
ST.DEV	0.716	0.734	0.583	0.659	0.758	0.704	0.572	0.607
TOTERR	0.723	0.737	0.585	0.659	0.758	0.704	0.573	0.608

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.054	-0.055	-0.033	-0.061	-0.088	-0.088	-0.033	-0.081
ST.DEV	0.742	0.744	0.579	0.609	0.723	0.575	0.579	0.544
TOTERR	0.744	0.746	0.579	0.612	0.729	0.582	0.579	0.550

BIAS - 2 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	-0.025	-0.023	0.012	-0.009	-0.019	-0.017	-0.004	0.056
ST.DEV	0.851	0.864	0.634	0.829	0.955	0.966	0.735	0.689
TOTERR	0.851	0.864	0.634	0.829	0.955	0.966	0.735	0.691

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.100	0.096	0.129	0.062	-0.004	0.026	0.126	0.049
ST.DEV	0.718	0.752	0.573	0.652	0.785	0.735	0.570	0.599
TOTERR	0.725	0.756	0.588	0.655	0.785	0.735	0.574	0.601

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.114	0.093	0.063	0.107	0.178	0.079	0.055	0.113
ST.DEV	0.818	0.837	0.621	0.752	0.890	0.797	0.621	0.595
TOTERR	0.826	0.842	0.624	0.761	0.907	0.801	0.623	0.598

BIAS - 3 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.156	0.164	0.088	0.066	0.078	0.083	0.063	0.175
ST.DEV	0.713	0.761	0.623	0.647	0.740	0.737	0.604	0.571
TOTERR	0.730	0.779	0.630	0.650	0.745	0.742	0.608	0.597

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.041	0.057	0.056	-0.011	-0.033	-0.058	0.041	0.095
ST.DEV	0.829	0.864	0.677	0.770	0.917	0.862	0.697	0.685
TOTERR	0.830	0.865	0.679	0.770	0.918	0.863	0.698	0.691

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.078	0.080	0.121	0.038	-0.029	0.028	0.121	0.059
ST.DEV	0.862	0.859	0.736	0.760	0.815	0.715	0.736	0.606
TOTERR	0.865	0.863	0.746	0.761	0.816	0.715	0.746	0.609

BIAS - 4 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.234	0.149	0.195	0.131	0.044	0.038	0.146	0.082
ST.DEV	1.062	1.091	0.871	0.916	0.968	0.965	0.882	0.663
TOTERR	1.088	1.092	0.893	0.926	0.969	0.966	0.894	0.666

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.304	0.181	0.220	0.134	0.138	0.142	0.177	-0.023
ST.DEV	0.913	0.972	0.721	0.827	0.991	0.924	0.697	0.622
TOTERR	0.962	0.989	0.754	0.838	1.001	0.935	0.719	0.622

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.139	0.097	0.162	0.126	0.073	0.026	0.143	0.100
ST.DEV	0.891	0.898	0.646	0.686	0.854	0.728	0.638	0.610
TOTERR	0.902	0.905	0.667	0.698	0.858	0.729	0.654	0.618

CONTAMINATION RATE - 10 PERCENT

BIAS - 1 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.073	0.068	0.050	0.023	0.053	0.054	0.049	0.078
ST.DEV	0.794	0.791	0.616	0.684	0.834	0.828	0.649	0.683
TOTERR	0.797	0.794	0.618	0.684	0.835	0.830	0.650	0.688

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.111	0.140	0.061	0.088	0.095	0.067	0.061	0.054
ST.DEV	0.807	0.801	0.659	0.741	0.879	0.844	0.659	0.658
TOTERR	0.815	0.813	0.662	0.747	0.884	0.847	0.662	0.660

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.096	0.088	0.056	0.049	0.099	0.056	0.060	0.052
ST.DEV	0.761	0.777	0.614	0.716	0.847	0.709	0.615	0.605
TOTERR	0.767	0.782	0.616	0.720	0.853	0.711	0.618	0.607

BIAS - 2 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.114	0.126	0.178	0.091	0.074	0.074	0.115	0.149
ST.DEV	0.889	0.907	0.730	0.863	1.041	1.025	0.794	0.610
TOTERR	0.897	0.995	0.752	0.868	1.043	1.027	0.803	0.629

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.242	0.266	0.228	0.293	0.325	0.264	0.218	0.150
ST.DEV	0.825	0.790	0.638	0.774	0.949	0.886	0.657	0.702
TOTERR	0.860	0.854	0.678	0.828	1.004	0.925	0.692	0.718

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.075	0.081	0.148	0.108	0.097	0.124	0.148	0.154
ST.DEV	0.763	0.764	0.651	0.709	0.841	0.742	0.651	0.685
TOTERR	0.767	0.768	0.668	0.717	0.847	0.753	0.668	0.702

BIAS - 3 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.408	0.366	0.429	0.321	0.309	0.318	0.347	-0.058
ST.DEV	1.046	1.023	0.898	1.033	1.174	1.169	0.992	0.691
TOTERR	1.124	1.085	0.996	1.083	1.214	1.211	1.051	0.694

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.262	0.198	0.256	0.119	0.072	0.096	0.219	0.189
ST.DEV	0.876	0.951	0.680	0.713	0.862	0.811	0.685	0.759
TOTERR	0.915	0.965	0.727	0.723	0.865	0.817	0.720	0.783

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.281	0.250	0.307	0.240	0.175	0.235	0.307	0.093
ST.DEV	0.921	0.898	0.696	0.788	0.927	0.783	0.696	0.631
TOTERR	0.963	0.932	0.761	0.824	0.943	0.817	0.761	0.638

BIAS - 4 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.405	0.298	0.333	0.143	0.116	0.116	0.171	0.053
ST.DEV	1.128	1.069	0.891	0.858	0.969	0.960	0.810	0.694
TOTERR	1.199	1.110	0.951	0.851	0.976	0.967	0.828	0.696

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.363	0.402	0.450	0.346	0.336	0.334	0.392	0.169
ST.DEV	1.151	1.100	1.012	0.988	1.152	1.149	0.990	0.696
TOTERR	1.208	1.228	1.109	1.048	1.201	1.197	1.065	0.717

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.441	0.313	0.421	0.301	0.216	0.152	0.400	0.200
ST.DEV	1.002	0.953	0.950	0.987	1.096	1.012	0.956	0.752
TOTERR	1.096	1.003	1.040	1.032	1.117	1.023	1.039	0.759

CONTAMINATION RATE - 15 PERCENT

BIAS - 1 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.161	0.248	0.181	0.181	0.166	0.147	0.172	0.211
ST.DEV	0.757	0.802	0.605	0.697	0.825	0.817	0.639	0.629
TOTERR	0.779	0.855	0.632	0.721	0.842	0.830	0.662	0.663

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.248	0.251	0.296	0.282	0.232	0.247	0.296	0.294
ST.DEV	0.759	0.776	0.638	0.726	0.868	0.821	0.652	0.659
TOTERR	0.799	0.846	0.704	0.779	0.899	0.857	0.716	0.722

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.168	0.129	0.136	0.160	0.186	0.184	0.136	0.249
ST.DEV	0.741	0.756	0.618	0.694	0.834	0.674	0.618	0.585
TOTERR	0.749	0.767	0.633	0.713	0.854	0.699	0.633	0.636

BIAS - 2 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.331	0.336	0.312	0.306	0.296	0.277	0.304	0.175
ST.DEV	0.937	0.976	0.778	0.891	1.028	1.026	0.836	0.734
TOTERR	0.954	1.002	0.839	0.943	1.070	1.063	0.890	0.755

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.203	0.147	0.257	0.222	0.186	0.175	0.240	0.239
ST.DEV	0.816	0.866	0.619	0.732	0.896	0.841	0.643	0.661
TOTERR	0.841	0.900	0.670	0.765	0.915	0.859	0.687	0.703

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.396	0.367	0.398	0.332	0.236	0.298	0.398	0.198
ST.DEV	0.965	0.997	0.804	1.000	1.164	1.072	0.804	0.705
TOTERR	1.044	1.068	0.898	1.054	1.188	1.113	0.898	0.731

BIAS - 3 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.262	0.246	0.389	0.279	0.234	0.233	0.291	0.268
ST.DEV	1.116	1.169	0.921	1.056	1.215	1.216	0.996	0.814
TOTERR	1.147	1.194	1.001	1.093	1.238	1.238	1.038	0.856

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.514	0.495	0.710	0.549	0.512	0.499	0.638	0.304
ST.DEV	1.105	1.165	0.934	1.022	1.246	1.224	0.953	0.849
TOTERR	1.266	1.287	1.175	1.161	1.348	1.323	1.148	0.902

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.542	0.565	0.495	0.425	0.408	0.368	0.598	0.314
ST.DEV	1.054	1.173	0.932	1.088	1.125	1.044	0.959	0.859
TOTERR	1.218	1.287	1.056	1.095	1.197	1.108	1.065	0.897

BIAS - 4 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.557	0.548	0.824	0.537	0.367	0.381	0.584	0.327
ST.DEV	1.245	1.399	1.140	1.286	1.429	1.428	1.243	1.002
TOTERR	1.366	1.503	1.409	1.395	1.476	1.478	1.374	1.055

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.568	0.355	0.556	0.276	0.190	0.208	0.403	0.140
ST.DEV	1.136	1.126	0.935	1.008	1.160	1.144	0.940	0.888
TOTERR	1.271	1.161	1.089	1.046	1.176	1.163	1.023	0.900

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.757	0.568	0.709	0.402	0.258	0.319	0.627	0.137
ST.DEV	1.346	1.458	1.063	1.262	1.494	1.462	1.095	0.774
TOTERR	1.546	1.547	1.280	1.326	1.516	1.497	1.263	0.786

CONTAMINATION RATE - 20 PERCENT

BIAS - 1 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.207	0.219	0.220	0.188	0.200	0.203	0.187	0.220
ST.DEV	0.729	0.796	0.568	0.659	0.816	0.810	0.802	0.615
TOTERR	0.758	0.826	0.609	0.686	0.840	0.835	0.831	0.653

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.203	0.276	0.207	0.266	0.283	0.303	0.211	0.229
ST.DEV	0.723	0.736	0.615	0.689	0.844	0.768	0.806	0.605
TOTERR	0.751	0.767	0.650	0.739	0.890	0.826	0.842	0.647

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.156	0.252	0.169	0.148	0.132	0.249	0.172	0.197
ST.DEV	0.755	0.782	0.604	0.772	0.919	0.734	0.802	0.628
TOTERR	0.771	0.816	0.627	0.786	0.929	0.776	0.827	0.659

BIAS - 2 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.465	0.393	0.487	0.458	0.430	0.427	0.456	0.296
ST.DEV	0.881	0.968	0.664	0.866	1.018	1.016	0.804	0.815
TOTERR	0.997	1.045	0.825	0.981	1.106	1.103	0.925	0.868

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.474	0.432	0.430	0.361	0.342	0.360	0.412	0.360
ST.DEV	0.847	0.851	0.688	0.768	0.911	0.874	0.896	0.716
TOTERR	0.972	0.955	0.812	0.849	0.974	0.946	0.810	0.803

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.356	0.356	0.371	0.222	0.138	0.130	0.366	0.261
ST.DEV	0.943	0.955	0.744	0.815	0.985	0.914	0.798	0.700
TOTERR	1.065	1.043	0.852	0.845	0.995	0.923	0.826	0.729

BIAS - 3 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.465	0.393	0.487	0.458	0.430	0.427	0.456	0.296
ST.DEV	0.881	0.968	0.664	0.866	1.018	1.016	0.804	0.815
TOTERR	0.997	1.045	0.825	0.981	1.106	1.103	0.925	0.868

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.474	0.432	0.430	0.361	0.342	0.360	0.412	0.360
ST.DEV	0.847	0.851	0.688	0.768	0.911	0.874	0.696	0.716
TOTERR	0.972	0.955	0.812	0.849	0.974	0.946	0.810	0.803

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.356	0.336	0.371	0.222	0.138	0.130	0.366	0.201
ST.DEV	0.943	0.955	0.744	0.815	0.985	0.914	0.740	0.700
TOTERR	1.009	1.013	0.832	0.845	0.995	0.923	0.826	0.729

BIAS - 3 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.453	0.278	0.502	0.307	0.240	0.249	0.348	0.386
ST.DEV	1.106	1.123	0.927	0.972	1.117	1.109	0.925	0.870
TOTERR	1.197	1.197	1.056	1.020	1.143	1.137	0.985	0.952

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.601	0.561	0.498	0.348	0.228	0.238	0.428	0.362
ST.DEV	0.957	1.042	0.780	0.804	0.929	0.909	0.771	0.809
TOTERR	1.132	1.139	0.927	0.876	0.956	0.940	0.883	0.887

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.575	0.571	0.515	0.479	0.439	0.454	0.517	0.296
ST.DEV	1.091	1.147	0.916	1.007	1.234	1.200	0.936	0.820
TOTERR	1.235	1.262	1.052	1.116	1.311	1.283	1.070	0.872

BIAS - 4 STANDARD DEVIATIONS

FILTER SIZE 1 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.589	0.459	0.706	0.426	0.266	0.263	0.485	0.536
ST.DEV	1.060	1.173	1.003	1.016	1.111	1.112	0.973	1.074
TOTERR	1.214	1.254	1.228	1.104	1.143	1.144	1.088	1.201

FILTER SIZE 2 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.941	0.745	0.886	0.665	0.531	0.531	0.729	0.364
ST.DEV	1.346	1.407	1.128	1.284	1.462	1.456	1.181	1.035
TOTERR	1.645	1.594	1.437	1.448	1.556	1.550	1.390	1.098

FILTER SIZE 3 STANDARD DEVIATIONS

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
MEAN	0.935	0.701	0.753	0.494	0.369	0.394	0.686	0.497
ST.DEV	1.268	1.207	0.938	1.036	1.230	1.199	0.943	1.107
TOTERR	1.578	1.466	1.205	1.148	1.284	1.263	1.164	1.215

3 Evaluation of a Machine/Method

3.1 Introduction

In this chapter, we deal with the evaluation of a single measuring process. By evaluation, we mean both statistical and physical assessments. The emphasis will be on the evaluation of autoanalysers, although much of what is said will also be pertinent to other measuring processes. Before we embark on this discussion, we describe what we mean by a measuring process.

3.2 Description of a Clinical Measuring Process

Let us identify the factors influencing measurements, i.e. the reported clinical results, which have been requested by the General Practitioner, Specialist or Ward Doctor. Consider the schematic representation of the request and result procedure shown in figure 3.1. We shall consider this diagram again in chapter 5 when considering intra-laboratory quality control. In this chapter, we are interested in that part which is contained in the large box labelled 'laboratory', and which is specified by 'Laboratory Measurement'.

We note that there are many factors which affect the clinical results. These factors are identified in the rectangles at the bottom of the diagram. The specific factors which affect the clinical results can be grouped under the headings: the measuring method, the reagents, the environment, the handling of the clinical results (both numerical and administrative), the calibration of the machine/method

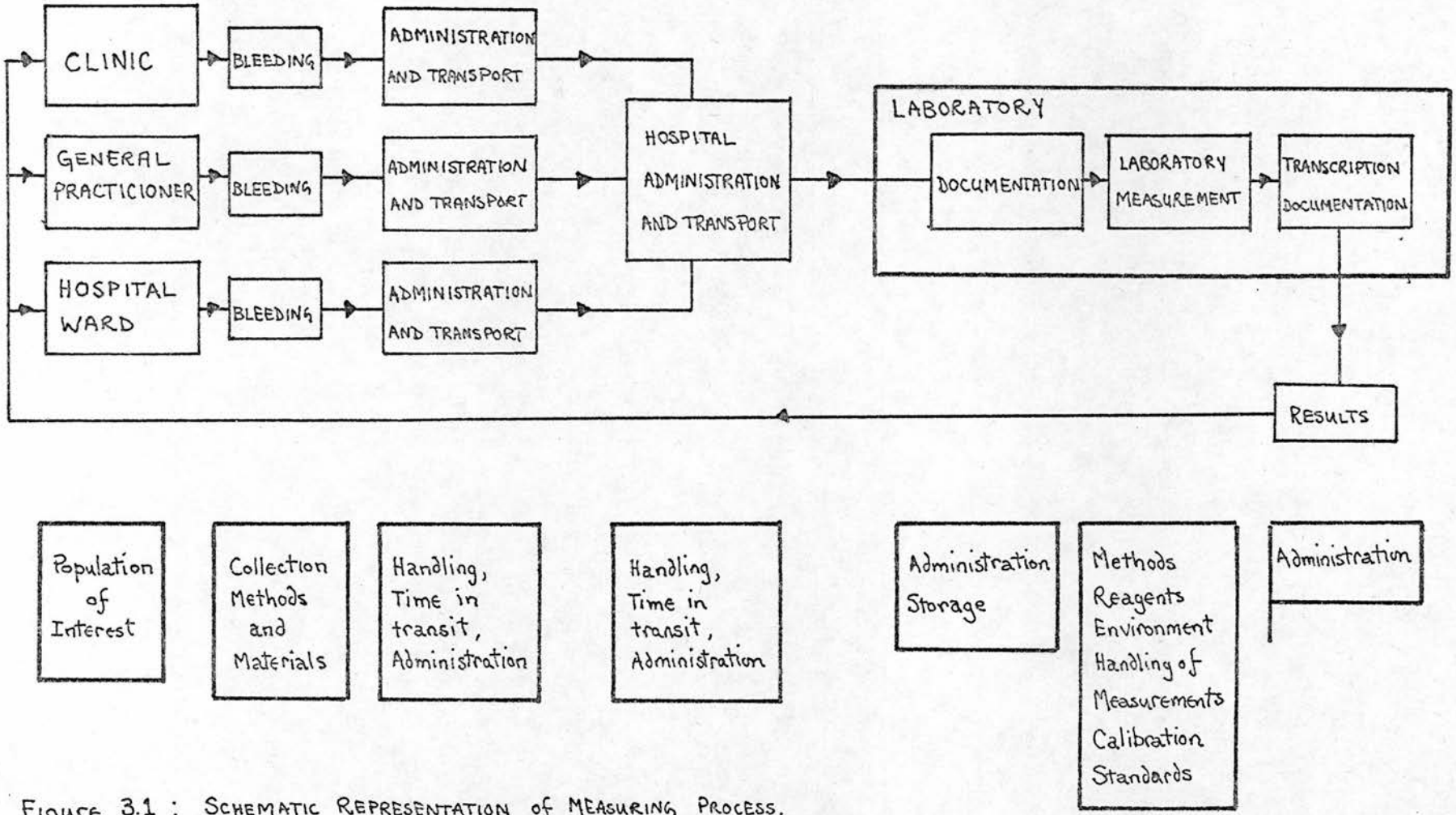


Figure 3.1 : Schematic Representation of Measuring Process.

and the standards which are used.

We shall deal specifically with measurements made with the use of machines, and in particular with machines such as the Coulter Counter Model S and Model S Senior. As we have indicated there are two components to the evaluation of a machine. In the following section we describe, briefly, the physical evaluation before discussing in depth the statistical evaluation in the sections 3.5ff.

3.3. Physical Evaluation.

The physical evaluation of a machine/method covers many facets of the hardware and the routine operation of the measuring process. A preliminary investigation of the general capabilities and characteristics of the machine will have taken place before any detailed investigations proceed. This initial study will have determined very roughly, whether the machine/method is capable of the required work, workload etc. Any machine not passing this initial check will not be considered for further investigation.

A more detailed list of points to investigate should be constructed after this brief study. The general form of such a list can be obtained from Sharpe (1970) or Broughton, Buttolph, Gowenlock (1969) and Broughton, Gowenlock, McCormack, Neill (1974). To paraphrase their major suggestions we give the following lists.

Points on basic information.

1. Basic cost.
2. Basic facilities and options for 'extras'.
3. Principle of measurement.
4. Rate of analysis. (Claim and reality).
5. Through-put time per sample.

6. Standards provided.
7. Space and service. (floor/bench space, drainage, power, water/gas input etc.)
8. Trained staff for routine service and/or maintainance.
9. Cost of, and arrangement of repair and maintainance. (Cost of contracts.)

Points on sample information.

1. Exact detail of sample acceptable. (e.g. haematology - anticoagulant, concentration; clinical chemistry - whole blood, separated plasma, serum, centrifuged blood.)
2. Facilities for rejection of unsuitable samples.
3. Size and presentation of batches to machine.
4. Limitation of specimen volume.

Points on reagent information.

1. Reagents required; supplied or made?
2. Details of preparation if reagents have to be made.
3. Volumes of reagents used per day, per hour.
4. Requirements for mixing of reagents and sample.
5. Reagents specifically prohibited.

Instrumentation.

Points to be considered under this heading are specific to the actual machine being investigated. For example, when considering cell counters we should have information on packed cell volume estimation, modes for output and the storage of results, the wavelength of colorimetry etc.

Miscellaneous Points.

1. Mode of results. (Does output fit medical records?)
2. Special presentation of specimens - cooled, incubated, heated etc.

3. Automatic sample identification.

These lists relate specifically to large auto-analysers, but most points made can be applied to the general area of machine/method evaluation.

In the following sections we discuss the statistical evaluation of such machine/methods. But, before we proceed, we should mention two other facets of machine evaluation which are sometimes overlooked.

The first of these is a measure of the reliability of the machine/method. The discussion of reliability cannot be meaningful without a discussion of accuracy, precision etc. - see 3.5. An accepted definition of reliability is

'the probability of the machine performing its purpose adequately for the period of time intended, under the operating characteristics encountered'.

For a discussion of the ramifications of such a definition see Calabro (B6) or Barlow and Proschan (B4). As we see the definition specifies that the machine must fulfil its intended job adequately. This can really only be judged after deciding exactly how we can measure whether a machine is or is not performing adequately.

The second facet concerns the interaction of man and machine. The interfacing of humans and machine obviously affects the measuring process. Technicians can be biased towards or away from accepting a machine for many reasons. Sharpe (1970) introduced the term Anxiety Factor to describe the anxiety to the operator when using complicated equipment or methods. By using his experience and skills, the operator can usually produce reliable results from a machine. However, this may only be achieved after spending much of the time adjusting, cleaning, maintaining, replacing certain elements etc. This, for

Sharpe, is the Anxiety Factor. In other words, the average time spent per day in manipulating the machine to produce adequate results. If this time is large, Sharpe suggests that user fatigue and/or neurosis will be experienced.

This type of index goes some way to describe the man/machine interface and the problems that are experienced. Personal preference for a particular machine would be accurately described as a subjective evaluation. A discussion of this can be found in Broughton et al (1974). These points again acknowledge the difficulty of quantifying such features. Some machines for example produce sickening smells, are noisy, require frequent adjustment, involve the operator in a great deal of activity. While results could be acceptable in terms of a statistical evaluation, the actual working conditions of staff in producing such measurements could be intolerable. One difficulty in assessing such problems is that often the relatively short evaluation period is not sufficient for such problems to come to the notice of the technicians.

3.4 Description of Coulter Counter Model S and Model S Senior.

As we have previously pointed out peripheral blood counting experienced a 'revolution' on the introduction by Coulter, of cell counting by electrical methods. We shall discuss the evaluation of auto-analysers. In particular, we discuss Coulter Counting systems. Recently a new version, called the S Senior, has been released as an improvement of the standard Coulter Counter models. This new version has some modifications which we discuss later. For the present we describe the standard model which is in common laboratory usage at this time.

The machine prints, in digital form the Red cell count, RBC, the White cell count, WBC, the Haemoglobin concentration, Hb, the Mean cell volume, MCV, the Mean cell haemoglobin, MCH, and the Mean cell haemoglobin concentration MCHC. The haematocrit, Hct, is also computed. (It should be noted that in other systems the haematocrit, which is the Packed cell volume, PCV is a primary measure.) It is capable of handling 150 - 180 blood samples per hour. (These figures are acceptable only as a target. In practice they will not be attained and perhaps operating at such a through-put rate is neither beneficial to machine nor staff.)

The Coulter principle is used in cell counting. The blood being analysed passes through an orificial tube. At the orifice, a meter measures resistance fluctuation. The passage of a cell causes a change of resistance (voltage). These changes are recorded as a count, and also as a magnitude. The magnitudes reflect the volumes of the cells which are being counted. Coincidence corrections are automatically employed, with certain specimens. Hb is determined by a cyanmethhaemoglobin technique. The MCV is computed from the eryth rocyte size distribution and the mean threshold setting being used. The remaining parameters are computed from these measurements.

The counter has five identifiable modules.

(i). A dilutor unit.

(ii). A processor.

(iii). A power unit.

(iv). A pneumatic power supply.

(v). A printout unit.

The operation, by either experienced or inexperienced staff, can be outlined schematically as follows, see Figure 3.2.

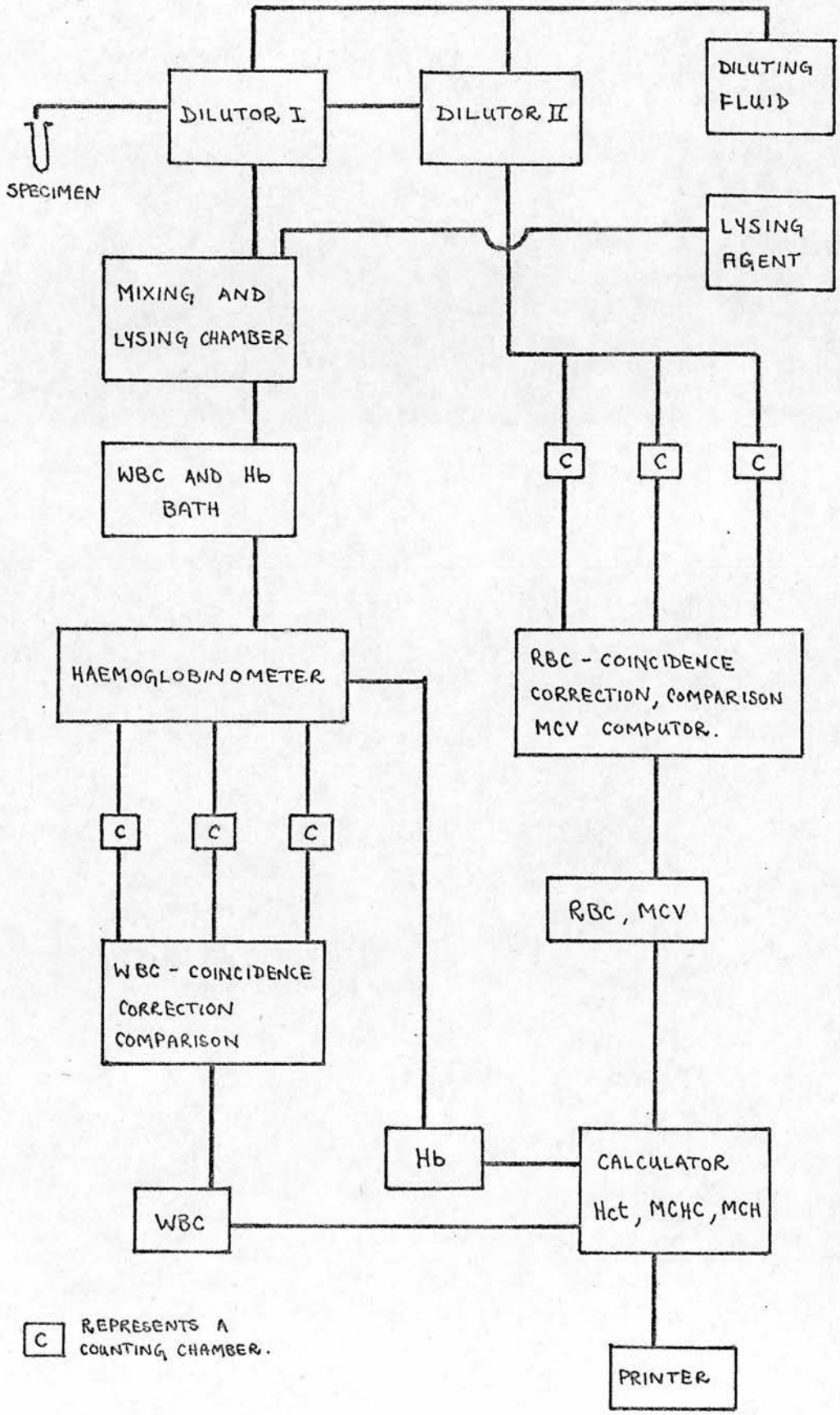


Figure 3.2 : Schematic Representation of Coulter Cell Counter Model S.

Specimens are presented individually to the intake. (Either of the intakes can be used - one for prediluted specimens, the other for whole blood) Approximately 1.3 - 1.4 ml of blood is aspirated. The bulk of this blood is used in flushing away the previous sample. The remaining 44.7 µl is used in the estimations described above. After dilution the volume of diluent and blood is split into two parts. One part is used in the estimation of RBC, while the other is used in the determination of Hb and WBC. In each of the estimations the counting is done in triplicate.

In sections 2.3.4-5 we discussed methods for handling triplicate measurements. Further, we showed how this particular system differed from the analysis of replicate measurements. The counters, can be independently calibrated. The rejection of results from these counters, or indeed of the specimen is an automatic procedure built into in Coulter-S. Warning lights indicate the particular counter(s) which are rejected. We remark at this stage that although we accept the necessity for a rejection procedure in each channel, the description of the procedure in the manual should be clearer and fuller.

The new version of the standard machine called the Coulter Counter Model S Senior as we have said, has some modifications. The most important one affecting our investigations is a change in the method of 'wash-out'. In the updated version, the wash-out is not only by flushing the system with a bulk of the incoming patient blood. A wash out fluid is also used. This as we show later will change the carry-over of the machine.

Having described the type of machine we are interested in evaluating we now turn our attention to the statistical evaluation of

such a machine. To do this we must define the indices which describe identifiable characteristics of the machine.

3.5 Definitions of Statistical Terms Concerning Evaluation.

We shall now define and investigate the innate characteristics of a measuring process. We restrict our attention to the Coulter-S, both standard and updated models, but most of our discussions are pertinent to other measuring processes. The identifiable and estimable sources of variation in a single measuring process will be called accuracy, precision, carryover, linearity and drift. In the introduction, we described briefly what we meant by accuracy and precision. We shall now define and discuss these terms together with the others mentioned in more detail.

3.5.1 Accuracy.

Suppose we estimate a certain clinical parameter for a specific patient as x. If we assume that the 'true' but unknown value is μ, then we could consider that accuracy is the degree of 'closeness' of x to μ. This is the basic ingredient of accuracy, but we must refine it. To see why this is necessary, let us consider making one measurement on the patient specimen with 'true' value μ. The difference between x and μ need not arise only due to an inaccuracy, it may be caused by a random error or indeed one of the other quantities described in this section. To surmount this 'one-off' problem, we consider the closeness of the measurements to 'true' values on average. Thus if we were to make replicate measurements on the same specimen, the average of these measurements, $\bar{x}$ would be compared with the 'true' value, to gauge accuracy. The closeness of this mean to the 'true' value can only be given meaning once we have

considered the precision of the measuring process.

The main problem with this definition of accuracy, is that it depends on the exact knowledge of the specimen's 'true' value. As we have previously stated, this knowledge is rare in haematological situations. There are some reference preparations available. This problem will be discussed more fully later, see Chapter 6.

3.5.2 Precision

The precision of a method/machine, is a measure of the degree of reproducibility of the process. Let us consider making replicate measurements on a single patient sample, which has a 'true' but unknown value μ. The measurements reported are $x_1, x_2, \ldots, x_n$. From the above, the accuracy is reflected in the quantity:

$$\mu - \bar{X}, \quad \text{where } \bar{X} = \frac{1}{n}\Sigma x_i \ .$$

The precision reflects the degree of spread in the measurements $x_1, x_2, \ldots, x_n$. That is the precision can be 'measured' by the variance (or standard deviation), defined by:

$$S^2 = \frac{1}{n-1} \Sigma(x_i - \bar{X})^2$$

We note that the larger this measure, the more imprecise the process is. (Some authors would refer to this measure as a measure of imprecision and not of precision. As long as we understand the measure the actual naming is one of personal taste.)

Now, theoretically, we could consider that precision should be measured by the index:

$$\hat{S}^2 = \frac{1}{n-1} \Sigma(x_i - \mu)^2$$

This quantity confounds accuracy and precision. It can be written in the following way:

$$\hat{S}^2 = \frac{1}{n-1} \Sigma(x_i - \bar{X})^2 + \frac{n}{n-1} (\bar{X} - \mu)^2$$

The first term is our measure of precision, i.e. the spread of the measurements, while the second term reflects the bias or inaccuracy of the process. It is necessary to construct indices which independently reflect the different process characteristics.

In our discussions we assume, unless otherwise stated, that the measuring process(es) has constant precision throughout the range of its clinical estimation. This is almost certainly unrealistic. It may well be that the precision is constant throughout the 'normal range'. Commonly, however, the precision decreases in the abnormal i.e. pathological ranges. Statistically, situations in which the precision varies leads to a complex if not impossible problem to solve. Some methods, relying on a weighting of measurements have been investigated. However, we suggest that the methods we give are reasonable approximations, giving answers which can be obtained by laboratory personnel.

3.5.3 Carryover

In very general terms, carryover is the degree of contamination between successive samples. Carryover is easily seen to be process dependent. For example, let us consider the Coulter-S standard model and also the modified version. As we have stated, the 'wash out' mechanism in the standard model relies basically on a part of the blood sample being used to flush the system. The modified version uses a washout fluid as well. Due to the differences in the viscosity of blood and wash-out fluid, and the associated turbulence we might well expect different carryover characteristics. We shall discuss this more fully later. We shall consider three procedures for the estimation of carryover, which have been reported in the literature. In fact, the method of estimation essentially defines carryover or

at least what the different authors mean by carryover,

(i). Pinkerton, Spence, Ogilvie, Ronald, Marchant and Ray (1970)

The device is flushed three times with isotonic 'particle free' solution and then,

(a) The isotonic solution is 'counted', d_1, and this is taken to be the 'background' count. This is subtracted from all subsequent counts.

(b) A blood sample is counted twice sequentially, results are a_1 and a_2. The first, a_1 is rejected but the second a_2 is recorded as the measurement.

(c) The isotonic solution is counted again, d_2.

(d) Carryover c, is determined as

$$c = \frac{d_2 - d_1}{a_2 - d_1}$$

The procedure is carried out ten times using a total of ten different bloods. The authors suggest that blood specimens with RBC ranging from 3.4 to 8.2 million/cmm be used in the scheme.

(ii). Broughton, Gowenlock, McCormack and Neill (1974).

In defining carryover the associated estimating procedure with reference to clinical chemistry, Broughton et al use a 'method of blanks'. In (i) above a similar idea was used. Such procedures seem to be fundamental in all commonly used procedures for the estimation of carryover. This procedure requires three successive aliquots of the same 'high valued' blood to be measured sequentially. These are reported as a_1, a_2, and a_3. These are

immediately followed by three aliquots of a 'low valued' blood, which give results b_1, b_2, and b_3. The carryover is estimated by k, which is:

$$K = \frac{b_1 - b_3}{a_3 - b_3}$$

This method, as with (i) above, assumes negligible drift. Further it assumes that the 'true' values of the bloods can be obtained from measurements made after the 'blanking' procedure, i.e. that a_3 and b_3 are in fact the true values of the specimens. It is not at all clear in this paper exactly how many of these determinations i.e. k, should be made. Neither is it clear exactly how we should combine these separate estimates to obtain an overall estimate of carryover. One of the problems which has been encountered with this estimate of carryover, is the frequency with which it yields a negative estimate of carryover; such values have no practical meaning. In fact unless estimators of carryover are constrained in some way, they will almost certainly give negative estimates on some occasions.

A negative estimate will occur when $a_3 - b_3$ and $b_1 - b_3$ have different signs. This may be produced by the difference in the true clinical values of the materials being small in comparison with the precision of the channel, the rounding of the reported values, negligible drift etc.

In practice we would infer from a negative estimate of carryover that the actual carryover is zero. However,

this is dependent on the assumption that the estimator of carryover is reasonable and has acceptable statistical properties.

A further estimator of carryover has been proposed by Broughton and Dawson (1970) for use when the precision of the 'channel' i.e. measuring process is poor.

(iii). A criticism of such procedures has been given by Hirst and Howorth (1975). For continuous channel analysis with a sample A followed by identical samples, both of B, the signal picture is shown in Fig. 3.3

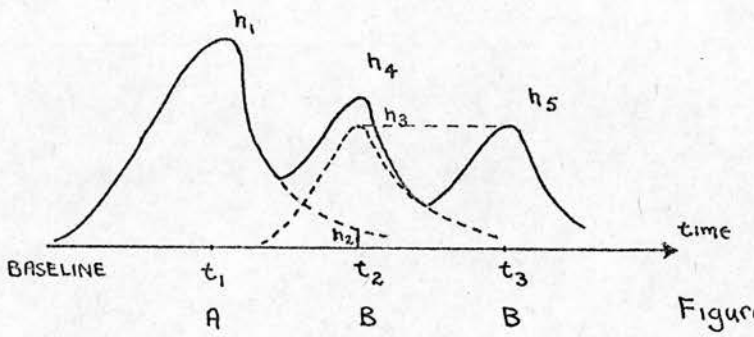


Figure 3.3

Now from (ii) above we would estimate carryover by

$$K = \frac{h_4 - h_3}{h_1 - h_3}$$

The suggestion made by Hirst and Howorth is that carryover will add a fixed proportion k' of the preceding peak where

$$K' = e^{-p/b}.$$

In this expression, p is the time interval between peaks, and b is estimated for the particular channel. We assume that the signal decays exponentially. To estimate, we set up a reagent 'baseline', then the carryover $h_2 = h_4 - h_5$ and the percentage carryover is

$$k'' = \frac{h_2}{h_1} \times 100 = \frac{h_4 - h_5}{h_1} \times 100$$

Hirst and Howorth comment that k is of little value, since it will vary throughout a run of random specimens, and further it may be infinitely large or even negative. The carryover defined by Hirst and Howorth is clearly more applicable to systems in which a signal analysis is carried out in order to ascertain the particular measurement. This is clearly more appropriate to machines such as the SMAC clinical chemistry auto-analyser than devices such as the Coulter-S.

(iv). Daniel (1974).

This description of carryover, discussed more fully later, is defined within the framework of a particular statistical model. It will be shown that under certain additional assumptions, that this definition reduces to that in (i) when only a diluent/blood sequence is investigated and to that in (ii) when a blood/blood sequence is investigated.

We note that the estimation of carryover must correspond to the type of contamination which is present in the system in routine laboratory use. In the standard model of the Coulter-S, the contamination is caused by the immediately preceding blood specimen. However, in the modified model, a wash-out procedure is used. The contamination is different, being in this case mainly caused by the contamination of the immediately preceding wash-out.

A further note should be made. The estimation of carryover is important since a large carryover effect will decrease the quality of the laboratory's work. It is especially important in situations where in routine conditions, the laboratory handles many varied pathological specimens. In such cases, the carryover between for example, a polycythaemic patient's specimen and an anaemic patient's specimen could affect the clinical parameters being measured.

3.5.4 Drift

Measuring machines, powered and controlled by complex electrical circuitry usually experience some drift during routine operation. By drift we mean the change in measurements made on the same specimen with respect to time, which cannot be attributed to carryover or a change in machine precision. Thus, if we were to measure the same blood specimen repeatedly throughout the day, and consider that there is neither carryover nor drift, then the values being reported by the machine should be

$$y_t = \mu + \varepsilon_t$$

where y_t is the measurement at time t, μ is the 'true' but unknown value of the specimen and ε_t is a random error. In the presence of drift, however, the situation could be represented by,

$$y_t = \mu + f(t) + \varepsilon_t$$

where f(t) is a term which is dependent on time only.

It is common to assume a linear or quadratic time effect. In most situations, however, this must be considered as an approximation to some more general function - perhaps sinusoidal. In many cases, it is held that drift should be estimated. We propose that we only

require to estimate drift in order to correct our results, rather than trying to estimate a particular functional form for drift. The particular functional form will depend on the current environmental conditions and other factors.

The problems of estimating drift, have been discussed by Youden (1954), and many others. We shall only note a few important points here. Drift is usually investigated by making a series of repeated measurements on the sample specimen. These repeated measurements should be independent (see 2.2). Youden has proposed the use of a statistic whose distribution was investigated by Hart (1942). Let a series of measurements, $y_1, y_2, \ldots, y_n$ be made on the same blood. Calculate

$$S^2 = \Sigma (y_i - \bar{y})^2 \quad \text{and} \quad D^2 = \Sigma d_i^2$$

where d_i is the difference between successive measurements. If no drift (trend) is present, the statistic D^2/S^2 should be less than 2. This type of analysis has been used by Pinkerton et al (1970). If however, we require to estimate drift, carryover etc. simultaneously this method cannot be used.

3.5.5 Linearity

We mean by this, the linearity of the specific machine. Let Ψ_t be the 'true' but unknown value of the t^{th} specimen. Due to carryover, the machine acutally measures a specimen with true value $g(\Psi_t)$. We shall discuss the functional form for g(.) later. The machine, however, acutally measures $f(g(\Psi_t))$ instead of $g(\Psi_t)$. What we mean by linearity is that we require f(.) to be a linear function. i.e. $f(z) = a + bz$.

3.6 Comments on Previous Inyestigations

The estimation of drift, carryover, linearity, accuracy and precision have in the past been done in what would appear to be an ad hoc basis. Usually each quantity has been estimated in separate experiments, set up in such a way that other influencing factors have been controlled to some degree. Apart from the problem of interdependence of these quantities, the amount of work being carried out to determine each is substantial. Consider Broughton's estimation of carryover by the method of blanks. Ten 'high' to 'low' and ten 'low' to 'high' sequences if used, require 63 measurements to be made, only 41 of these measurements are actually used in the estimation of carryover. Further, no correction is made for other factors e.g. drift. We shall give a procedure which firstly utilises all collected information, and secondly acknowledges the interdependence of the quantities.

A more 'sophisticated' procedure was suggested by Daniel (1974). Basically we will use a similar approach later. Consider for the present Daniel's model without the drift corrections. He suggests a model

$$y_t = f(X_t) + \varepsilon_t$$

where $f(X_t)$ is a quadratic function in x_t and ε_t is a random error. This would apply in situations with no carryover. However, each diluted sample loses a fraction θ, of it's original volume, V, in passing through tubes etc., and picks up a volume θV of material from the walls, be it washout or previous blood sample. The fraction θ, is assumed to be identical for every sample being analysed. This may not be exactly true, but will be a reasonable assumption to make.

With this formulation, the actual value to be measured is

$$X'_t = \frac{(1-\theta)VX_t + \theta VX'_{t-1}}{(1-\theta)V + \theta V} = (1-\theta)X_t + \theta X'_{t-1} \qquad 3.6\ (i)$$

We note that this implies a contamination by samples $X_{t-2}, X_{t-3}, \ldots$, since

$$X'_{t-1} = (1-\theta)X_{t-1} + \theta X'_{t-2}$$

so that

$$X'_t = (1-\theta)X_t + \theta(1-\theta)X_{t-1} + \theta^2 X'_{t-2}$$

In fact we could write this in the form

$$X'_t = (1-\theta) \sum_{o}^{t-2} \theta^k X_{t-k} + \theta^{t-1} X_1$$

An obvious approximation is appropriate at this stage, namely, that since carryover will usually be not more than 10% and usually less than 5%, we can neglect terms in θ^2 and higher powers of θ.

Allowing for a quadratic measuring process and a quadratic time effect, Daniel proposes the following model:

$$y_t = b_o + b_1 X'_t + b_{11} X'^2_t + d_1 t + d_{11} t^2 + \varepsilon_t \qquad 3.6\ (ii)$$

We use 3.6 (i) and note that $X'_{t-1} \approx X_{t-1}$, since we have stated that carryover is less than 10% and usually less than 5%. So 3.6(ii) reduces to:

$$y_t = b_o + b_1\left[(1-\theta)X_t + \theta X_{t-1}\right] + b_{11}\left[(1-\theta)X_t + \theta X_{t-1}\right]^2 + d_1 t + d_{11} t^2 + \varepsilon_t$$

$$y_t = b_o + b_1(1-\theta)X_t + b_1\theta X_{t-1} + b_{11}(1-\theta)^2 X_t^2 + 2b_{11}(1-\theta)\theta X_t X_{t-1} + b_{11}\theta^2 X_{t-1}^2 + d_1 t + d_{11} t^2 + \varepsilon_t \qquad 3.6\ (iii)$$

Daniel has found, and it would appear to be reasonable from other experience, that terms in θ^2 and X_{t-1}^2 and $X_t X_{t-1}$ are negligible. Thus equation 3.6 (iii) reduces to:

$$y_t = b_o + b_1(1-\theta)X_t + b_1\theta X_{t-1} + b_{11}(1-\theta)^2 X_t^2 + d_1 t + d_{11} t^2 + \varepsilon_t$$

This was the model used by Daniel. The basic problem experienced by Daniel was that no design could be found enabling the orthogonal estimation of the various parameters. Daniel used five equally spaced specimen levels, judging each design by the efficiency of estimation. We find even more difficulty, since it will not usually happen that we can obtain specimens which are equally spaced. Daniel gives only approximations to the parameters standard errors.

If we consider Daniel's model without the drift and x_t^2 term we have:

$$y_t = b_o + b_1(1-\theta)X_t + b_1\theta X_{t-1} + \varepsilon_t \qquad 3.6\ (iv)$$

In our present notation, the previous estimates of carryover see 3.5.3(ii) have been given by the index

$$K = \frac{y_t - X_t}{X_{t-1} - X_t} \qquad 3.6\ (v)$$

where X_t is the 'true' value of blood B and X_{t-1} the 'true' value of blood A. The time factor is not important, since we assume no drift. The measured, contaminated, value of blood B is y_t. If we re-arrange the equation 3.6 (v) we have:

$$y_t = X_t + K(X_{t-1} - X_t)$$

or $$y_t = (1-K)X_t + KX_{t-1} \qquad 3.6\ (vi)$$

If we compare this with equation 3.6 (iv), we see that implicit in the previous estimation of carryover are

(a) No bias in calibration. (i.e. $b_o = 0$)

(b) No proportional deviation (i.e. $b_1 = 1$).

(c) A different error structure from Daniel's additive proposal.

The error structure proposed by Daniel allows for independent, homoscedastic errors on each measurement but in the cases described

before, we have no clear statement of the error assumptions. It seems that in each of these cases it is implied that

$$K = \frac{b_1 - b_3}{a_3 - b_3} + \varepsilon \qquad \text{(see 3.5.3 (ii))}$$

However, it is perhaps more likely that an appeal is made to the Central Limit Theorem in any inference statement to be made. (It should be noted that in each of the previous estimations, we have a ratio of dependent random variables - in each case the numerator and denominator are measurements made by the process.)

3.7 Proposed General Model.

We shall introduce a somewhat ideal model, which will be subsequently modified. Let $y_t, y_{t-1}, y_{t-2}, \ldots, y_1, y_0$ be the actual measurements recorded by the machine. Let $\Psi_t, \Psi_{t-1}, \Psi_{t-2}, \ldots, \Psi_1, \Psi_0$ be the true but unknown values of the corresponding specimens. Let $X_t, X_{t-1}, \ldots, X_1, X_0$ be the values of the samples which the machine actually measures. These are different from the $\Psi_t, \Psi_{t-1}, \ldots, \Psi_0$ due to the contamination of the preceding specimen.

We can write

$$X_t = g(\theta, \Psi_i : i = 0,1,2,3,\ldots,t) \qquad 3.7 \text{ (i)}$$

The measuring process of the machine produces results for these X_k, $k = 0,1,2,\ldots,t$, of $f(X_k)$, $k = 0,1,2,\ldots,t$. The function $f(.)$ need not necessarily be linear. These measurements are made at times t, $t = 0,1,2,\ldots$ i.e. equally spaced instances. The drift at these times can be written as $d(t)$, $t = 0,1,2,\ldots$. If we assume additivity we have:

$$y_t = f(X_t) + d(t) + \varepsilon_t \qquad 3.7 \text{ (ii)}$$

where ε_t is a random error. We shall assume later that these errors have a normal distribution, but this may not be a reasonable assumption in some situations. The full model in detail is

$$y_t = f(g(\theta, \Psi_i : i = 0,1,2,\ldots t)) + d(t) + \varepsilon_t \qquad 3.7 \text{ (iii)}$$

Assumptions and modifications.

Assumptions concerning drift : d(t).

As we have stated previously, the actual functional form of d(.) could realistically have any form. For the design of experiments, the actual time taken to run the experiment is short. The estimation of drift throughout a day or week or longer is very much more complicated problem. We are not interested in the estimation of d(.), the functional form. Presumably at the initial evaluation stage, a machine which oscillated or trended wildly on repeated measurements would have been rejected as unsuitable for normal laboratory use. We are interested in using our estimate of drift as a correction term, allowing us to estimate the other parameters of interest more efficiently. Since the time interval over which our experiment will run is short, we propose to consider a linear drift. This appears to be a reasonable assumption in the light of experience. This will allow for orthogonal estimation. Even in the presence of quadratic or higher order drift effect the approximation by a linear term is expected to be good. (see figure 3.4)

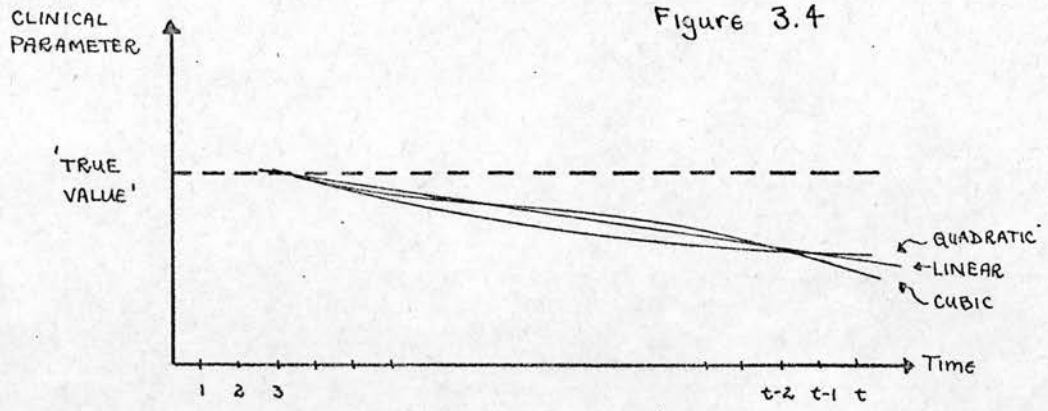


Figure 3.4

Assumptions concerning carryover: g(.) - the contamination function.

Our previous proposal was that the contamination function was such that we could write:

$$X_t = g(\theta, \Psi_i : i = 0,1,2,\ldots,t)$$

We denote by θ the carryover, and consider the mathematical formulation in much the same approach as Daniel. We have seen that by considering a proportion θ of the last specimen being gained by the current sample, while θ of the current sample is lost we had, in our present notation:

$$X_t = (1-\theta)\Psi_t + \theta X_{t-1}$$

This can be represented by

$$X_t = (1-\theta) \sum_{o}^{t-2} \theta^k \Psi_{t-k} + \theta^{t-1}\Psi_1 \qquad 3.7 \text{ (iv)}$$

As we have remarked before, θ will be much less than 1.00, usually between 0.02 and 0.06. Thus powers of θ will have little effect on our considerations, and so we neglect terms involving θ^2, θ^3 etc. Equation 3.7 (iv) reduces to

$$X_t = (1-\theta)\Psi_t + \theta\Psi_{t-1} \qquad 3.7 \text{ (v)}$$

This formulation states that samples are contaminated only by their most recent predecessor.

Assumptions concerning linearity; f(.)

The functional form of f(.), determines the basic measuring characteristics of the machine. If f(.) is linear, we would have in some sense an ideal machine.

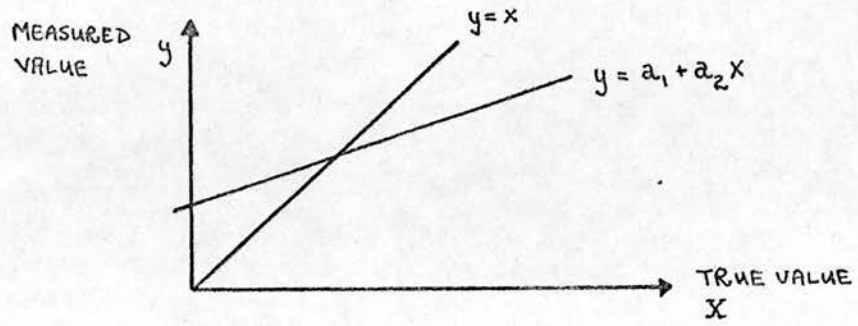


Figure 3.5

If we have, in Figure 3.5, $f(x) = a_1 + a_2x$, then a_1 is the bias (additive) in measurement, while a_2 is a proportional error if different from 1.00. In the ideal situation we would have $a_1 = 0$ and $a_2 = 1.00$ i.e. $y = x$.

We propose that the function f(.) is linear, and in some situations we can test this assumption. As a result of this, we have a model of the form

$$y_t = b_o + b_1(1-\theta)\Psi_t + b_1\theta\Psi_{t-1} + b_2t + \varepsilon_t \qquad 3.7\ (vi)$$

Assumption concerning the error structure: ε_i, i = 0,1,2,....

The model given in equation 3.7 (vi) is not linear, due to the terms $b_1(1-\theta)\Psi_t$ and $b_1\theta\Psi_{t-1}$. Further no transformation will reduce it to a linear model. We shall assume that the random errors are Gaussian distributed with zero mean and common variance σ^2. Further we assume that the errors are independent of the true blood value and of each other.

The distributional assumptions will be reasonable in this case, although not always. In some situations in the presence of a peculiar rounding process this may not be a reasonable assumption.

The assumption of independence of errors is not apparently so appealing. Firstly, due to the sequential nature of the process, it may be that the error are correlated in some way. Secondly, we have assumed independence between the error and the actual blood level e.g. no tendency towards small errors with bloods in the most sensitive region of measurement and larger errors in the extreme pathological regions. The necessary extensions to the present treatment will, hopefully be investigated elsewhere.

To use the model proposed, we require to construct a design for an experiment allowing for the efficient estimation of the parameters of interest. In the next section we deal with these design considerations.

3.8 Design Considerations.

We wish to estimate the parameters - $b_o, b_1, , b_2$ and σ^2, to evaluate the machine. However, we note that in order to estimate these parameters, the Ψ_t, $t = 0,1,\ldots,n$ i.e. the 'true' blood values must also be estimated, or at least be estimable. Thus we see that in a sequence of n laboratory measurements, we have n+5 parameters to estimate - given that all bloods in the sequence are different.

To overcome this problem - without recourse to external information e.g. precision of the machine, which we wish to estimate at this instant - we must use at least some of the blood specimens for replicate measurement. We could do this by randomly measuring blood specimens, but this would in no sense be efficient. We 'control' the values of Ψ_t, $t = 0,1,..,n$. That is, the true values of the blood specimens being measured are not known quantitatively, but

they are known to be replications of one of the bloods being used. The different fixed, known or unknown values of Ψ_t will be called levels.

We shall now investigate how we might construct sequences of blood specimens, which enable estimation of the various parameters with efficiency. Further, we also require that such estimation is effected in a reasonably short sequence of bloods.

The first question to answer is, 'how many levels should be used for such a sequence'? If carryover must be estimated, at least two levels must be used. If linearity must also be investigated, at least three levels must be used. We shall assume that a minimum of three levels will be used.

Having decided this, we consider the following constraints on the design (sequence)

(i). The design must be balanced for time effects.

(ii). All levels must appear an equal number of times.

(iii). All sequences of 'pairs' must appear an equal number of times.

The first constraint (i) seems intuitively obvious for efficient estimation. The second constraint (ii), requires the same number of measurements to be made on each of the blood specimen levels used. This type of constraint is commonly found in designing experiments - in chapter 4 we mention this when considering completely balanced designs. It can be shown that such designs are more efficient than designs which have the same total number of measurements, but which have unequal number of measurements (replications) for each level. Thus if N measurements are made in the design (sequence), we require r replicates for each of the l levels used. Hence we note that $N = r.l$.

The third constraint also appears to be reasonable in the light of our assumptions concerning carryover. We have assumed that contamination due to carryover is present in the current sample only from it's immediate predecessor.

Let us for the present, consider the construction of designs using three levels. That is, we use three different bloods. The number of 'pairs' in a sequence of N is N-1. As an example if we have three different bloods represented by A, B, and C used in the sequence

A B C

we see that there are two pairs, AB and BC. We note that the other possible pairs AC, BA, CB, and CA, AA, BB, CC do not appear. It should be noticed that the constraints (ii) and (iii) are not compatible. We therefore replace the constraint (ii) by the constraint (iia).

(iia). A design of size N+1 is used, in which the first level being 'measured' is used only to provide the 'missing' pair.

Thus if three levels are used, we note that nine pairs of levels must appear - AA, BB, CC, AB, AC, BA, BC, CA, CB. Thus we would require $Q = 9x+1$ measurements, where x is the number of times a 'pair' of levels appear. Q is the total number of measurements which the machine makes and N is the number of these measurements actually used in the analysis.

From these constraints, we require both $Q = 9x+1$ and $Q = 3r+1$ where r and x are both integer. Obviously we only need to consider $Q = 9x+1$ as the other is immediately satisfied if a design can be constructed with this constraint.

For the moment, we shall consider the use of three levels. These levels will be denoted H, M, and L which can be thought as representing bloods which are in the clinical High, Medium and Low ranges. For most auto-analysers there is a constant though-put time for specimens, at least when the machines are operating correctly. Thus we indicate the position of the different specimens in the sequence by what may be considered as time codes - 0,1,...3r.

Minimal Design Construction.

To construct a minimal design, i.e. the smallest sequence of specimens needed to satisfy the constraints (i), (iia) and (iii), we remember that if Q is the 'size' of the sequence, then Q = 9x+1 for x integer. Thus, we need only consider sequences of the following sizes, Q = 10, 19, 28, 37, ..., that is N = 9, 18, 27, ... Let us consider each of these in turn, until we can construct at least one design (sequence) which satisfies the constraints given.

N = 9: This, as we have previously shown, is the minimum size of design which would allow for a balance in level 'pairs'.

A possible sequence with this property is

H : L L M M H H M L H

All pairs appear once and all levels appear three times, with the exception of H. Consider the balance in time. We can give the above sequence the following time codes

0 : 1 2 3 4 5 6 7 8 9

H : L L M M H H M L H

For such a sequence to be balanced with respect to time, we require the sum of the time codes for each level to be the same. In this case we see that the sums for H, M and L are

20, 14 and 11 respectively.

Can any sequence with N = 9 which is balanced for pairs also be balanced for time? Essentially, we require to partition the numbers 1,2,3,...,9 into three disjoint sets, each containing three numbers, two of which are in sequence. The sum of each of these sets must be 15(1+2+3+...+9 = 45). If we consider the set which contains 9, we must have the other two time codes summing to 6. Further, these two must either be in sequence or the sequence must start with this level at 0 and 1 and the level must appear at the time code 5. We can immediately rule out the first of these suggestions since we cannot use the time codes 8 and 9 - these already have a sum greater than 15 - nor can we have two codes in sequence which have a sum of 6. Let us suppose that we have given the level, M, the time code 9. The remaining possibility we can represent as follows.

0 : 1 2 3 4 5 6 7 8 9

M : M . . . M . . . M

Thus, we must fill the position 2,3,4,5,6,7, and 8 with three H's and three L's, two of each appearing in sequence. Further, we must achieve this in such a way that the sums of the codes for these is 15. This can be done only in the following ways:

0 : 1 2 3 4 5 6 7 8 9

M : M H L L M H H L M

or M : M L H H M L L H M

Neither of these satisfy our constraints concerning the appearance of 'pairs'. For the first sequence the pairs MH,

HL and LM all appear twice to the exclusion of the pairs HM, LH and ML. For the second sequence, the pairs HM, LH and ML appear twice to the exclusion of the pairs MH, HL and LM. Thus we can reject the suggestion that a design can be constructed for N = 9. The argument given is general and does not rely on a specific level appearing first.

N = 18: In this case, each level is replicated six times. Each of the 'pairs' of levels must appear exactly twice. Thus, for viable designs, we must at least satisfy this constraint. Consider the level H. Each of the following pairs must occur:

HH HM HL MH LH

Only one sequence of HHH or two sequences of HH will satisfy the condition that HH appears twice only.

Consider the case of a single sequence HHH. To ensure that the remaining H pairs can be present, there must be three isolated H's. By doing this, the isolated H's can supply six of the outstanding H pairs, while the triple sequence supplies 4. Thus the 10 H pairs can be obtained.

Consider the case of two sequences of HH. This configuration gives 6 possible H pairs, and so we require the remaining 2 H's to be isolated. It will be obvious that there are no other configurations which can furnish a possible balance for pairs of levels.

As examples of such sequences, we have

(a) M : H M M M L L L H H H L H L M L M H M

(b) M : M H H L L H M L M M L M H L L H H M

We shall call the designs of the form (a), i.e. containing a triple sequence of each level, a Triple Design. Designs of the form (b), i.e. containing two double sequences of each level will be called Double designs. We note that the designs given in (a) and (b) above satisfy the criteria which we have laid down i.e. time balance, pair balance and each level appears an equal number of times - excepting the first specimen whose value is not used.

In the following sections, we investigate the properties of these designs. Later in this chapter, we extend the discussion to cover the construction of such sequences in general, extending the area to include designs for more than three levels.

3.9 The Use of Coding

As we have previously stated, the use of coding when considering time is an obvious step. The machine makes measurements in a sequential procedure with relatively constant intervals between determinations. (Obviously a breakdown or stall in the measuring procedure may invalidate this). We shall code time in a more convenient way than before. Let us denote by 1,2,3,...,18 the number of the specimen being measured, the corresponding time codes being -17, -15, -13,...-1, 1, 3, ..., 15, 17.

In some situations it may be possible to code the blood levels. As we have stated interest lies in estimating precision, carryover linearity and accuracy. In practice as we shall see the assessment of linearity is difficult in the haematological situation. Further the bias of the measuring process is confounded with the 'true' values of the clinical material used. Thus, the term b_o in the model 3.7 (vi)

can be given an arbitrary value which, in the following discussion we set equal to the mean of the levels.

If we knew that the blood levels of the clinical parameter being measured were such that the M value was exactly half way between the H and L values we could code the blood levels by -1, 0 and 1. This situation may arise if materials could be mixed, or perhaps diluted. However, it should be noted that while in other disciplines this situation can occur e.g. clinical chemistry, problems with the mixing of different blood specimens and the clinical parameters which we wish to measure, will usually prevent us from using such coding in peripheral blood counting.

Investigation of Triple and Double designs for Coded Values.

We consider in this section, situations in which both time and levels can be coded. The general model has been given, equation 3.7 (vi) and is:

$$y_t = b_o + b_1(1-\theta)\Psi_t + b_1\theta\Psi_{t-1} + b_2 t' + \varepsilon_t \qquad t' = (-19+2t)$$
$$t = 1,2,..,18$$

We can write this in a slightly different form, by reparameterising.

$$y_t = \beta_o + \beta_1\Psi_t + \beta_2\Psi_{t-1} + \beta_3 t' + \varepsilon_t \qquad t' = (-19+2t)$$
$$t = 1,...,18.$$

The matrix formulation of this is:

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

where $\underset{\sim}{y}$ is a vector of observations. (18x1)

X is the 'incidence' matrix corresponding to the levels, of -1, 0, and 1's. (18x4)

$\underset{\sim}{\beta}$ is a vector of parameters (4 x 1)

$\underset{\sim}{\varepsilon}$ is an error vector. (18 x 1)

TRIPLE DESIGN It will be shown later that at least one other Triple design exists, which satisfies the constraints discussed in section 3.8. The specific design which we consider was given previously as design (a) and is

M : H M M M L L L H H H L H L M L M H M

The specific representation for X in this case is:

$$X = \begin{bmatrix} 1 & 1 & 0 & -17 \\ 1 & 0 & 1 & -15 \\ 1 & 0 & 0 & -13 \\ 1 & 0 & 0 & -11 \\ 1 & -1 & 0 & -9 \\ 1 & -1 & -1 & -7 \\ 1 & -1 & -1 & -5 \\ 1 & 1 & -1 & -3 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 3 \\ 1 & 1 & -1 & 5 \\ 1 & -1 & 1 & 7 \\ 1 & 0 & -1 & 9 \\ 1 & -1 & 0 & 11 \\ 1 & 0 & -1 & 13 \\ 1 & 1 & 0 & 15 \\ 1 & 0 & 1 & 17 \end{bmatrix}$$

Then we see that

$$X'X = \begin{bmatrix} 18 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 \\ 0 & 0 & 12 & 0 \\ 0 & 0 & 0 & 1938 \end{bmatrix}$$

DOUBLE DESIGNS It will be shown later, that twenty four double designs can be constructed. Each of these designs satisfy the constraints which we have discussed. Statistically these sequences are equivalent as we see later. We shall consider a specific sequence of levels which has been used by Daniel. The sequence is:

M : M H H L L H M L M M L M H L L H H M

From what has gone above we see that this leads to:

$$X = \begin{bmatrix} 1 & 0 & 0 & -17 \\ 1 & 1 & 0 & -15 \\ 1 & 1 & 1 & -13 \\ 1 & -1 & 1 & -11 \\ 1 & -1 & -1 & -9 \\ 1 & 1 & -1 & -7 \\ 1 & 0 & 1 & -5 \\ 1 & -1 & 0 & -3 \\ 1 & 0 & -1 & -1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 3 \\ 1 & 0 & -1 & 5 \\ 1 & 1 & 0 & 7 \\ 1 & -1 & 1 & 9 \\ 1 & -1 & -1 & 11 \\ 1 & 1 & -1 & 13 \\ 1 & 1 & 1 & 15 \\ 1 & 0 & 1 & 17 \end{bmatrix}$$

This leads to

$$X'X = \begin{bmatrix} 18 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 \\ 0 & 0 & 12 & 0 \\ 0 & 0 & 0 & 1938 \end{bmatrix}$$

The form of the matrix X in these cases is similar i.e. only the middle columns are different and that only in internal arrangement. The form of the matrix X'X is identical in both types of sequence. The matrix is called diagonal. This implies that the parameters β, β_1, β_2 and β_3 can be orthogonally estimated. This condition is desirable in general and certainly we would dismiss any design which did not have this property - since we have been able to construct an orthogonal design.

Now we discuss how we might estimate the parameters using either of the designs which have been given.

3.10 Estimation of Parameters.

We have stated, when describing the general model and its formulation, that the 'usual' assumptions concerning the error structure will be made. These assumptions are listed below.

(i). The errors ε_i, $i = 1,2,\ldots,18$ have identical Gaussian distributions, with mean 0 and variance σ^2. i.e.
$\varepsilon_i \sim N(0,\sigma^2)$ $i = 1,2,\ldots,18$

(ii). The errors are independent of each other, i.e.
$\text{Cov}(\varepsilon_i,\varepsilon_j) = 0$ $i \neq j$

(iii). The errors are independent of the levels used in the design. i.e. $\text{Cov}(\varepsilon_i,\Psi_t) = 0$ $\forall i,t$.

The standard 'least squares' theory, states that the 'best' estimators of the parameters, $\underset{\sim}{\beta}$ are given by $\hat{\underset{\sim}{\beta}}$, where

$$\hat{\underset{\sim}{\beta}} = (X'X)^{-1} X'\underset{\sim}{y}.$$

Using this result, we investigate the estimation of the parameters in each of the design sequences which we have discussed. Estimation is carried out in the other possible designs - (see the remaining 23 Double Designs section 3.13.2) - in a similar manner.

We restate the reparameterised model.

$$y_t = \beta_o + \beta_1 \Psi_t + \beta_2 \Psi_{t-1} + \beta_3 t' + \varepsilon_t \quad t' = -19+2t,\ t = 1,2,..,18$$

The least squares estimators of the parameters are given in table 3.1 for the two sequences of levels i.e. Triple and Double designs.

TABLE 3.1

PARAMETER ESTIMATOR	Triple	Double
$\hat{\beta}_o$	$\frac{1}{18} \Sigma\, y_t$	$\frac{1}{18} \Sigma\, y_t$
$\hat{\beta}_1$	$\frac{1}{12}[(y_1+y_8+y_9+y_{10}+y_{12}+y_{17}) - (y_5+y_6+y_7+y_{11}+y_{13}+y_{15})]$	$\frac{1}{12}[(y_2+y_3+y_6+y_{13}+y_{16}+y_{17}) - (y_4+y_5+y_8+y_{11}+y_{14}+y_{15})]$
$\hat{\beta}_2$	$\frac{1}{12}[(y_2+y_9+y_{10}+y_{11}+y_{13}+y_{18}) - (y_6+y_7+y_8+y_{12}+y_{14}+y_{16})]$	$\frac{1}{12}[(y_3+y_4+y_7+y_{14}+y_{17}+y_{18}) - (y_5+y_6+y_9+y_{12}+y_{15}+y_{16})]$
β_3	$\frac{1}{1938} \Sigma\, t'y_t$	$\frac{1}{1938} \Sigma\, t'y_t$

Due to our reparameterisation, we have, in both cases, the following equations.

$$\hat{\beta}_o = \hat{b}_o$$
$$\hat{\beta}_1 = \widehat{b_1(1-\theta)}$$
$$\hat{\beta}_2 = \widehat{b_1\theta}$$
$$\hat{\beta}_3 = \hat{b}_2$$

From these equations, we see that for the original model we have constructed estimators for b_o and b_2. Estimators can easily be constructed for θ and b_1 from the above equations. Let these estimators be denoted by $\hat{\hat{\theta}}$ and $\hat{\hat{b}}_1$. Thus,

$$\hat{\hat{b}}_1 = \hat{\beta}_1 + \hat{\beta}_2$$

$$\hat{\hat{\theta}} = \frac{\hat{\beta}_2}{\hat{\beta}_1 + \hat{\beta}_2}$$

To write these estimators in terms of the experimental observations, the following sums are defined for each design.

TRIPLE:

$$S_{t1} = (y_1+y_{14}+y_{16}+y_{17})-(y_2+y_5+y_{15}+y_{18})+2(y_8+y_{12})-2(y_{11}+y_{13})$$

$$S_{t2} = (y_5+y_6+y_7+y_{11}+y_{13}+y_{15})-(y_1+y_8+y_9+y_{10}+y_{12}+y_{17})$$

DOUBLE:

$$S_{d1} = y_2+y_3+y_6+y_{13}+y_{16}+y_{17} - (y_4+y_5+y_8+y_{11}+y_{14}+y_{15})$$

$$S_{d2} = y_3+y_4+y_7+y_{14}+y_{17}+y_{18} - (y_5+y_6+y_9+y_{12}+y_{15}+y_{16})$$

Then the estimators can be written as follows.

TABLE 3.2

PARAMETERS	Triple	Double*
$\hat{b}_o$	$\frac{1}{18} \Sigma\, y_t$	$\frac{1}{18} \Sigma\, y_t$
$\hat{\hat{b}}_1$	$-\frac{1}{12} (S_{t1}+2S_{t2})$	$\frac{1}{12} (S_{d1}+S_{d2})$
$\hat{\hat{\theta}}$	$\frac{S_{t1} + S_{t2}}{S_{t1} + 2S_{t2}}$	$\frac{S_{d2}}{S_{d1} - S_{d2}}$
$\hat{b}_2$	$\frac{1}{1938} \Sigma\, t'y_t$	$\frac{1}{1938} \Sigma\, t'y_t$

(*For any of the other double designs, the counterparts to the sums S_{d1} and S_{d2} are

S_{d_1} = Sum of measurements on high specimens - Sum of measurement on Low specimens

S_{d_2} = Sum of measurements following High Specimens - Sum of measurements following Low specimens.)

It can be shown that the estimators $\hat{b}_o$, $\hat{b}_1$, $\hat{\theta}$ and $\hat{b}_2$ are the least squares estimators obtained by minimising the sum of squares

$$(\underset{\sim}{y} - \underset{\sim}{X\beta})' \ (\underset{\sim}{y} - \underset{\sim}{X\beta})$$

directly in the usual way with respect to b_o, b_1, θ and b_2.

Having constructed estimators for the parameters of interest in the original model, we now determine the distributional properties of these estimators. This will enable statements concerning confidence intervals to be made or hypothesis testing to be carried out.

Estimation of $\beta_o(b_o)$.

In both Triple and Double designs, we estimate β_o by the mean of all observations. It can be easily shown that in both cases the estimators are unbiased and have variance $\frac{\sigma^2}{18}$. Under the assumptions we have made the estimators have Gaussian distributions.

i.e. $\hat{b}_o \sim N(b_o, \frac{\sigma^2}{18})$

Estimation of b_2

In both Triple and Double designs, we estimate b_2 by

$$\hat{b}_2 = \frac{1}{1938} \Sigma\, t'y_t = \frac{1}{1938} \Sigma\, (2t-19)y_t$$

Again it can be shown that $\hat{b}_2$ is an unbiased estimate of b_2 and, for both designs, has a variance $\frac{\sigma^2}{1938}$. Under the assumptions we have made $\hat{b}_2$ has a Gaussian distribution.

i.e. $\hat{b}_2 \sim N(b_2, \frac{\sigma^2}{1938})$

Estimation of σ^2

In both Triple and Double designs, the estimate of precision, denoted $\hat{\sigma}^2$ is obtained from basic least squares theory. The unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{SSE}{n-(k+1)}$$

where $SSE = \underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{\hat{\beta}}'\underset{\sim}{X}'\underset{\sim}{y}$, and n is the number of observations made and k the number of parameters estimated. In our particular situation, n = 18 and k = 4.

From the assumptions made and basic statistical theory, $\frac{\hat{\sigma}^2}{\sigma^2}$ has a Chi-square distribution with 13 degrees of freedom.

Estimation of b_1

We have noted that the estimators for b_1 are

$$\text{Triple : } \hat{\hat{b}}_1 = -\frac{1}{12}(S_{t1} + 2S_{t2})$$

$$\text{Double : } \hat{\hat{b}}_1 = \frac{1}{12}(S_{d1} - S_{d2})$$

Both estimators can be shown to be unbiased and both have variance $\frac{\sigma^2}{6}$. Under the assumptions which have been made both estimators have Gaussian distributions.

$$\text{i.e. } \hat{\hat{b}}_1 \sim N(b_1, \frac{\sigma^2}{6})$$

Estimation of θ

The estimators which have been constructed for θ are ratios of dependent variables. Let us denote by $\hat{\hat{\theta}}_t$ and $\hat{\hat{\theta}}_d$ the estimators in the Triple and Double designs. Then these can be written as follows,

$$\hat{\hat{\theta}}_t = \frac{S_{t1} + S_{t2}}{S_{t1} + 2S_{t2}} \quad \text{and} \quad \hat{\hat{\theta}}_d = \frac{S_{d2}}{S_{d1} - S_{d2}}$$

We note that:

$$E(S_{t1} + S_{t2}) = -12b_1\theta \qquad E(S_{d2}) = 12b_1\theta$$

$$E(S_{t1} + 2S_{t2}) = -12b_1 \qquad E(S_{d1} - S_{d2})12b_1$$

$$V(S_{t1} + S_{t2}) = 12\sigma^2 \qquad V(S_{d2}) = 12\sigma^2$$

$$V(S_{t1} + 2S_{t2}) = 24\sigma^2 \qquad V(S_{d1} - S_{d2}) = 24\sigma^2$$

$$Cov(S_{t1}+S_{t2}, S_{t1}+2S_{t2}) = 12\sigma^2 \qquad Cov(S_{d2}, S_{d1}-S_{d2}) = 12\sigma^2$$

Further, in both cases the ratios are of dependent Gaussian variables. We require to investigate the ratio of two dependent Gaussian random variables.

Fieller (1932) has discussed this problem. The basic problem has subsequently been discussed by many other authors. A detailed discussion of the possible distributional shapes of the ratios of two dependent Gaussian random variables has been given by Marsagilia (1965). In some situations, the resulting distribution can be approximated well by a Gaussian distribution.

Commonly, the interest centres on the construction of confidence intervals for the ratio of two dependent Gaussian random variables. In the situation which has been discussed, namely the estimation of carry-over, this will almost certainly be the case. Fieller (1954) has given the construction procedure for such intervals. The essential features are presented in Kendall and Stuart (B23). We sketch the procedure briefly.

Let us investigate the ratio $\alpha = \frac{\eta}{\xi}$. We can rewrite this in the form

$$\eta - \alpha\xi = 0$$

Let x and y be unbiased estimates of ξ and η with variances and covariance V_{xx}, V_{yy} and V_{xy} respectively and further let us assume that x and y have a bivariate Gaussian distribution.

The fiducial range for $\frac{\eta}{\xi}$ is

$$(y^2-t^2V_{yy}) - 2\alpha(xy-t^2V_{xy}) + \alpha^2(x^2-t^2V_{xx}) \leq 0$$

We note that $\dfrac{y-\alpha X}{\sqrt{V_{yy}-2\alpha V_{xy}+\alpha^2 V_{xx}}}$ has a t-distribution.

The inequality can be considered as an inequality in α and t^2, and on re-arranging,

$$t^2 = \frac{y^2V_{xx}-2xyV_{xy}+x^2V_{yy}}{V_{xx}V_{yy}-V_{xy}^2} - \frac{\{(yV_{xy}-xV_{yy})-(yV_{xx}-xV_{xy})\alpha\}^2}{(V_{xx}V_{yy}-V_{xy}^2)(V_{yy}-2\alpha V_{xy}+\alpha^2V_{xx})}$$

Due to the inequality, we can reject certain values immediately. There is one asymptote to the curve, at $t^2 = \frac{x^2}{V_{xx}}$, corresponding to the probability level where $\frac{y}{x}$ ceases to be significant. The maximum value t^2_{max} is given by

$$t^2_{max} = \frac{x^2}{V_{xx}} + \frac{(yV_{xx}-xV_{xy})^2}{V_{xx}(V_{xx}V_{yy}-V_{xy}^2)}$$

which occurs at

$$\alpha_o = \frac{yV_{xy}-xV_{yy}}{yV_{xx}-xV_{xy}}$$

The upper and lower limits are given by α_1 and α_2, where

$$\alpha_1,\alpha_2 = \frac{(xy-t^2V_{xy}) \pm \{(xy-t^2V_{xy})^2 - (x^2-t^2V_{xx})(y^2-t^2V_{yy})\}^{\frac{1}{2}}}{x^2-t^2V_{xx}}$$

Tests for Linearity

For both designs, let $\bar{y}_L$, $\bar{y}_M$ and $\bar{y}_H$ be the mean values of the observations at Low, Medium and High levels (irrespective of the preceding level). From considerations of the models, we have

$$\bar{y}_L = b_o + b_1(1-\theta)L + b_1\theta(L+d) + \frac{1}{6}\sum_L \varepsilon$$

$$\bar{y}_M = b_o + b_1(1-\theta)M + b_1\theta(L+d) + \frac{1}{6}\sum_M \varepsilon$$

$$\bar{y}_H = b_o + b_1(1-\theta)H + b_1\theta(L+d) + \frac{1}{6}\sum_H \varepsilon$$

where $d = H-M = M-L$ and $\sum_L$, $\sum_M$ and $\sum_H$ denotes summation over those time positions for L, M and H respectively.

If we let A denote either L, M or H and we consider either the Triple or Double designs, we have

$$E(\bar{y}_A) = b_o + b_1(1-\theta)A + b_1\theta(L+d)$$

$$V(\bar{y}_A) = \frac{\sigma^2}{6}$$

$$\mathrm{Cov}[\bar{y}_A, \bar{y}_B] = 0 \quad A \neq B = L, M \text{ or } H$$

Consider the statistic

$$\eta = \bar{y}_H + \bar{y}_L - 2\bar{y}_M$$

Under the assumptions we have made, it can be shown that η has a normal distribution with zero mean and variance σ^2

i.e. $\eta \sim N(0,\sigma^2)$.

Thus, under the assumption of a linear measuring process, we see that the statistic, η, may be used as a test statistic. If we consider the alternative to be a quadratic i.e. we consider a model

$$y_t = b_o' + b_1'[(1-\theta)\Psi_t + \Psi_{t-1}] + b_3[(1-\theta)\Psi_t + \theta\Psi_{t-1}]^2 + b_2 t' + \varepsilon_t.$$

Under the same assumptions, but this model, η will have a distribution

$$\eta \sim N(2(1-\theta)^2 b_3' d^2, \sigma^2)$$

The approximate power of this test is (α = 5%)

$$\text{Power} = 1-\Phi\left(1.96+\frac{2(1-\theta)^2 b_3' d^2}{\sigma}\right) + \Phi\left(-1.96+\frac{2(1-\theta)^2 b_3' d^2}{\sigma}\right)$$

$$\text{where } \Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

3.11 Investigations of Triple and Double Designs for uncodable situations

In the previous section, we have considered situations in which the 'levels' being used can be coded i.e. they are known to be equidistant with respect to the clinical parameter being measured. Further, we have assumed that the time at which measurements are made can also be coded. In most situations, as we have indicated, the coding of time will be justified. However, the coding of levels, especially in haematological situations, is difficult to justify. It is difficult, if not impossible to realise a situation in which the levels being used can be 'manufactured' to be equidistant. In Clinical Chemistry when equidistant levels are required, two sera are mixed (pooled) to produce a third intermediate level. In this way, equidistant levels can be 'manufactured'. However, in haematological investigations using the Coulter Sprogramme attempts at such a procedure are seriously affected by agglutination and other interactions between different bloods. Further, even if this were to be achieved, we would require to carry out such a pooling procedure for each clinical parameter of interest, since we cannot achieve equidistance in all clinical parameters simultaneously.

Consider the model previously used.

$$y_t = b_o + b_1(1-\theta)\Psi_t + b_1\theta\Psi_{t-1} + b_2 t' + \varepsilon_t \qquad t' = 2t-19$$
$$t = 1,2,\ldots,18$$

Once more, we assume that the time positions can be coded. In this case, however, we designate the three levels appearing in Ψ_t and Ψ_{t-1} by H(high), M(medium), and L(low).

When we are able to code the levels, the parameters b_o and b_1 were required to describe the calibration of the device. However, when we introduce H, M and L, which require estimation, we no longer require b_o nor b_1. (In fact to include them would overparameterise the model. Essentially estimation of machine bias and estimation of 'true' levels are confounded.) The model in the uncodable situation reduces to the following.

$$y_t = (1-\theta)\Psi_t + \theta\Psi_{t-1} + b_2 t' + \epsilon_t \qquad \Psi_t = H,M,L, \quad t' = 2t-19$$
$$t = 1,2,\ldots,18.$$

The model is still non-linear, due to the terms in θH, θM and θL. However, we note that for θ fixed, the model is linear in the remaining parameters. To fit the model to the observed data, we could consider a number of different approaches.

(i). Proceed as before, employing least squares to obtain a set of equations whose solution(s) lead to estimates (i.e. solve a set of 'normal equations'). To solve the set of equations, initial values of the parameters will be required and an iterative procedure employed.

(ii). Initial values are given to the parameters θ,H,M,L and b_2 and a procedure for fitting a non-linear model is used. Such procedures can be found in computer packages e.g. Powell's method, Simplex method etc.

(iii). In practice, we can assume that θ will be within the interval (0.000 , 0.100) i.e. that carryover will be less than 10%. For a series of values of θ in this range, we evaluate the multiple regression for the remaining parameters, obtaining the residual sum of squares. From a list of the value of θ and the

corresponding sum of squares we can determine the value of θ which gives the minimum value of residual sum of squares.

(iv). A one dimensional search for θ is used, and at each value of θ selected, a standard multiple regression is carried out to estimate the remaining parameters. At each evaluation, the value of the residual sum of squares provides with previous evaluations, the direction in which the 'optimal' value of θ lies.

We choose the approach outlined in (iv). It should be remarked, however, that this approach is merely a slight sophistication of that described in (iii). In practice for a laboratory which has facilities for multiple regression analyses, but little else, approach (iii) will give, if used carefully, good estimates of all parameters required. (To do this, interpolation may be required if the intervals at which θ is set are large.) Further on average the number of computations required in (iii) is greater than in (iv).

We shall describe exactly the method of analysis in section 3.12.

3.12 Analysis of Practical Experiments.

We now discuss experiments which were carried out using the new modified version of the Coulter-S counter. These experiments were carried out in the Department of Haematology, Royal Infirmary Edinburgh. The presentation will contain a brief introduction covering the aims of the experiments, the experimental results, the statistical analysis and finally a discussion of the results with an indication of the proposals for use.

Aims.

(i). To investigate procedures for the simultaneous estimation of carryover, calibration, precision and drift for auto-analysers. (We note that investigation of linearity has not been specified. In the uncodable situation, difficulties in such an investigation will be seen).

(ii). To compare documented schemes for the estimation of carry-over with the schemes given in (i).

Materials.

The machine used in all experiments was the Coulter Counter Model S. Senior. The 'bloods' used were commercial preparations, manufactured by Coulter i.e. the name of the preparation is 4C, which is a registered preparation for use as a reference control, not as a standard or calibrator. The particular lots used were 1366/1, 3040/1 and 4033E/1.

Outline of Experiments.

(i) We propose to investigate the primary measuring channels of the Coulter -S i.e. WBC, RBC and Hb. The MCV channel was also investigated. The calibration (accuracy), carryover and precision were examined after corrections for possible drift.

The designs required three 'blood' levels. As we have stated, ideally these levels should be at the High, Medium and Low levels for all four clinical parameters. With patient bloods, this would be difficult to achieve. Further, we have also seen that the estimation procedure is simplified if it can be arranged that the level are such that,

High = Low + 2A and Medium = Low + A, where A is some constant.

We note that if, for some particular bloods this could be achieved in one clinical parameter it will not be true for the remaining three clinical parameters.

It was decided that commercial preparations be used. Coulter have recently manufactured 'reference controls' for Abnormal high and Abnormal low, as well as the frequently used 'normal' control. These are marketed under the name 4C.

We denote by H, abnormal high, M, normal and L, abnormal low. Both the Triple and Double designs were investigated. The sequence of measurements are given below.

	0		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
(a)	M	:	H	M	M	M	L	L	L	H	H	H	L	H	L	M	L	M	H	M
(b)	M	:	M	H	H	L	L	H	M	L	M	M	L	M	H	L	L	H	H	M

We note that the amount of blood required to carry out both (a) and (b) is approximately 16.8ml of H and L, and 18.2ml of M.

(ii) The sequence of measurements described in (i) above was followed by two other sequences of measurements. Both of these sequences have been used in practice to estimate carryover. Let a specimen of diluent be denoted by D. The same preparation of H and L were used. The laboratory's own 'normal' preparation was used and is denoted M'. The two sequences were:

(c)	D	D	D	H	H	H	L	L	L	D	D	D				
(d)	D	D	D	M'	M'	M'	M'	M'	M'	M'	M'	M'	M'	D	D	D

The machine was calibrated using the laboratory's own scheme, (see section 5.8.3) and was adjusted till it was considered to be in control. (For a discussion of the quality control scheme see chapter 5). Since the results and discussion for (ii) can be easily described, we shall deal with them first.

(ii)a The results are given in Table 3.3.

Table 3.3

	WBC(10^9/l)	RBC(10^{12}/l)	Hb(g/dl)	MCV(fl)
Diluent 1	0.2	0.02	0.1	1
Diluent 2	0.2	0.01	0.1	1
Diluent 3	0.2	0.01	0.1	1
H 4	21.4	6.14	20.0	92
H 5	22.0	6.25	20.4	92
H 6	-	5.94	20.2	93
M 7	3.9	3.76	9.5	72
M 8	3.5	3.71	9.5	72
M 9	3.4	3.68	9.4	71
Diluent 10	0.3	0.09	0.2	66
Diluent 11	0.3	0.02	0.1	1
Diluent 12	0.2	0.01	0.1	1

From these data we calculate three estimates of carryover per channel. Unlike the previous Coulter -S, we would expect the carryover from diluent to blood to be approximately the same as from blood to blood. The estimated carryover is given in table 3.4.

Table 3.4

CARRYOVER.	WBC(10^9/l)	RBC(10^{12}/l)	Hb(g/dl)	MCV(fl)
D → H	2.8%*	-3.4%	1.0%	1.1%
H → M	2.7%*	3.5%	1.0%	4.5%
M → D	3.1%	2.2%	1.1%	92.9%

These estimates of carryover were obtained by using Broughton et al's (1974) method i.e. if we have two, triplicate sequences of 'bloods'

$a_1\ a_2\ a_3\ b_1\ b_2\ b_3$

then carryover $= \dfrac{b_1 - b_3}{a_3 - b_3} \times 100$

We should note that entries marked '*' have been calculated using 'b_2' in place of 'b_3', since the machine rejected the WBC count in the third determination. Further as we had expected, the estimates of carryover for each channel are similar irrespective of sequence. However, as we will see later also, the carryover blood → diluent for the M.C.V. is uncharacteristically high. Lastly we should note that one of these estimates is negative. This bears out a previous criticism of this method of calculating carryover. (See section 3.5.3).

(ii)b The results are given in table 3.5.

Table 3.5

	WBC(10^9/l)	RBC(10^{12}/l)	Hb(g/dl)	MCU(fl)
Diluent 1	0.0	0.01	0.0	0
Diluent 2	0.1	0.02	0.0	0
Diluent 3	0.1	0.02	0.0	0
M' 4	3.8	4.62	14.8	98
M' 5	3.9	4.75	15.1	98
M' 6	4.1	4.79	15.1	98
M' 7	4.0	4.74	15.0	98
M' 8	3.9	4.81	15.0	98
M' 9	4.1	4.77	15.0	99
M' 10	4.6	4.68	14.8	98
M' 11	4.1	4.76	15.1	98
M' 12	4.0	4.78	15.2	98
M' 13	4.0	4.80	15.1	99
Diluent 14	0.2	0.12	0.3	83
Diluent 15	0.1	0.02	0.0	0
Diluent 16	0.1	0.03	0.0	0

Using these data, we calculate the carryoyer. The mean and standard deviation of the control bloods are also calculated.

It will be obvious that the calculation of accuracy and the standard deviation are used to investigate the accuracy and precision of the process. Since the measurements at 4 and 5 i.e. the first two measurements after the sequence of diluent, are obviously affected by carryover - i.e. are not true estimates of the value of M', we do not include them in the calculation of the mean nor the standard deviation. The estimates for accuracy, precision and carryover are given in table 3.6.

Table 3.6

	WBC(10^9/l)	RBC(10^{12}/l)	Hb(g/dl)	MCV(fl)
Mean, $\bar{X}$	4.10	4.77	15.04	98.25
Precision* (St.Dev.)	0.2138	0.0414	0.1188	0.4629
Carryover** (%)				
D→M'	7.5	3.2	1.6	0.3
M'→ D	2.5	1.9	2.0	84.5

* These entries calculated on n = 8 observations using

$$\sqrt{\frac{1}{n-1} \Sigma(x_i-\bar{x})^2}$$

** These entries calculated using Broughton et al (1979) formula

$$K = \frac{b_1 - b_3}{a_3 - b_3}$$

where b_3 and a_3 are replaced by $\bar{x}$ where appropriate.

All values for $\bar{x}$ and Precision calculated using readings (6-13) only.

Table 3.7 gives the Coulter values for the precision of each channel as stated in the reverse side of their pre-printed result card.

Table 3.7

	WBC(10^9/l)	RBC(10^{12}/l)	Hb(g/dl)	MCV(fl)
Coulter Precision	0.2	0.035	0.1	1.0

Comparing, the estimates of precision given in table 3.6 with the manufacturers values given in table 3.7, we see that there is close agreement.

From table 3.6 we note again the peculiarity exhibited by the MCV channel when carryover is estimated. The small changes experienced in measurements and the discretisation cause the estimates of carryover to be doubtful. We have seen this before - see table 3.4.

In chapter 5, we shall discuss the acceptable level of carryover for such peripheral blood counts. The level taken is commonly 3%. With such a value we note that, for example, the D-M' value for WBC is 'too large'. However, the main problem with, in this case, two estimates of carryover per channel is how one should combine them.

(i) In the following section, we present results of experiments using the two designs which we have discussed. Before doing this we briefly indicate how the analyses were carried out.

The computing was carried out on EMAS, Edinburgh Multi-Access System. The computer program, see Appendix 2, incorporates four NAG (Numerical Algorith Group) subroutines. The subroutine E02AAF performs a quadratic search for the minimum residual sum of squares, the subroutine GOZBDF computes the necessary information for the multiple

regression routine GO2CHF. Finally, the subroutine FO1ACF computes the inverse matrix which is the covariance matrix of parameter estimates. A flow-chart for this program is given in figure 3.6.

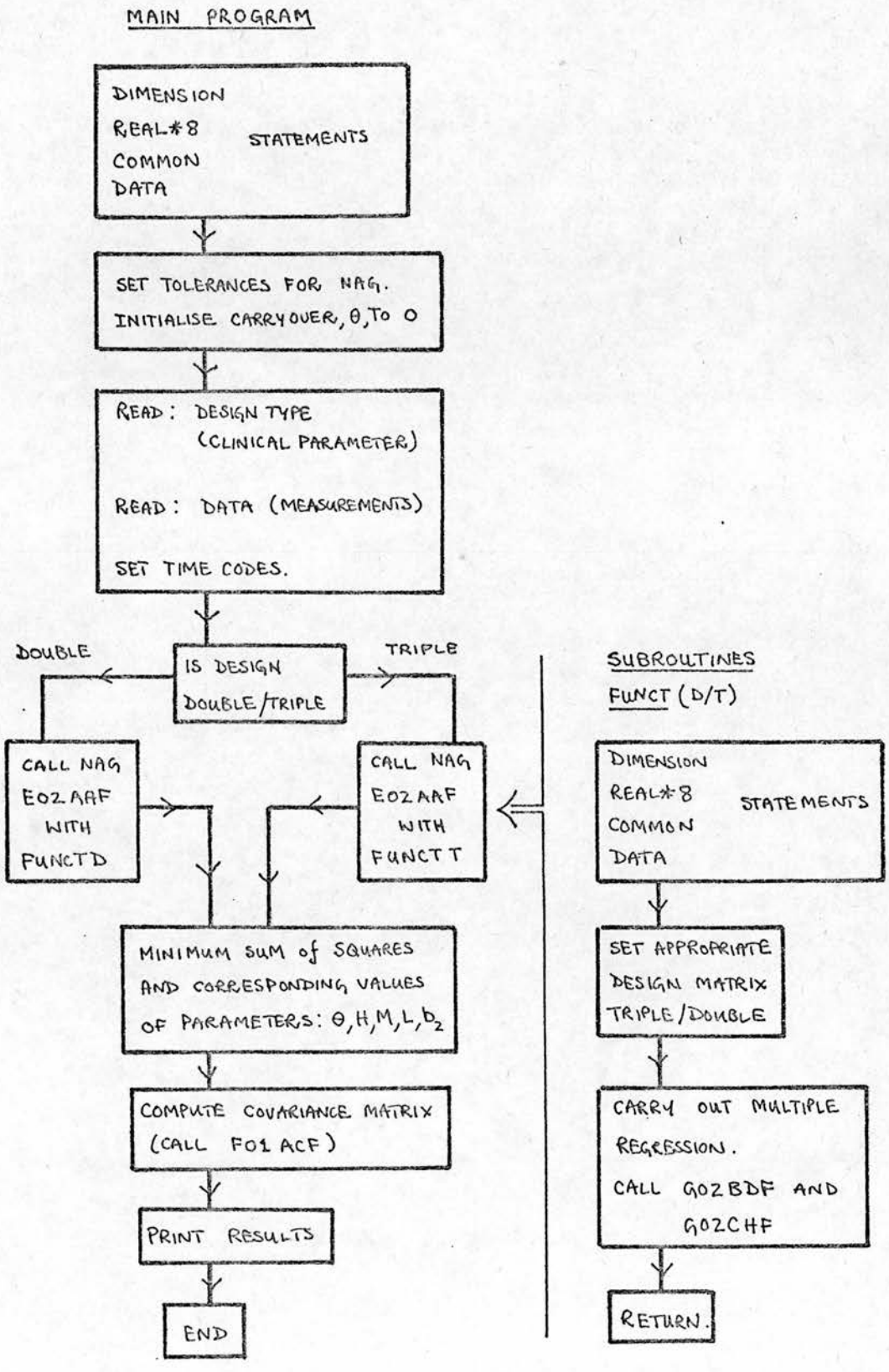

MAIN PROGRAM
DIMENSION
REAL*8
COMMON
DATA
STATEMENTS
SET TOLERANCES FOR NAG.
INITIALISE CARRYOVER, θ, TO 0
READ: DESIGN TYPE
(CLINICAL PARAMETER)
READ: DATA (MEASUREMENTS)
SET TIME CODES.
DOUBLE
IS DESIGN
DOUBLE/TRIPLE
TRIPLE
CALL NAG
E02AAF
WITH
FUNCTD
CALL NAG
E02AAF
WITH
FUNCTT
MINIMUM SUM of SQUARES
AND CORRESPONDING VALUES
OF PARAMETERS: θ, H, M, L, b2
COMPUTE COVARIANCE MATRIX
(CALL F01ACF)
PRINT RESULTS
END
SUBROUTINES
FUNCT (D/T)
DIMENSION
REAL*8
COMMON
DATA
STATEMENTS
SET APPROPRIATE
DESIGN MATRIX
TRIPLE/DOUBLE
CARRY OUT MULTIPLE
REGRESSION.
CALL G02BDF AND
G02CHF
RETURN.

Figure 3.6

As we have indicated previously, for a specified value of $\theta = \theta_o$ the model

$$y_t = (1-\theta_o)\Psi_t + \theta_o\Psi_{t-1} + b_2t' + \varepsilon_t$$

reduces to a linear model. The particular form which this takes is

$$y_t = H\, u(t,\theta_o) + M\, v(t,\theta_o) + L\, w(t,\theta_o) + b_2t' + \varepsilon_t.$$

The variables $u(t,\theta_o)$, $v(t,\theta_o)$ and $w(t,\theta_o)$ are dependent on the value of θ(i.e. θ_o), t and the design. For the designs, Triple and Double which we have used the values of these variables are presented in table 3.8.

Table 3.8

Triple :	$u(t,\theta_o)$	$v(t,\theta_o)$	$w(t,\theta_o)$	Double :	$u(t,\theta_o)$	$v(t,\theta_o)$	$w(t,\theta_o)$
t = 1	$1-\theta_o$	θ_o	0		0	1	0
2	θ_o	$1-\theta_o$	0		$1-\theta_o$	θ_o	0
3	0	1	0		1	0	0
4	0	1	0		θ_o	0	$1-\theta_o$
5	0	θ_o	$1-\theta_o$		0	0	1
6	0	0	1		$1-\theta_o$	0	θ_o
7	0	0	1		θ_o	$1-\theta_o$	0
8	$1-\theta_o$	0	θ_o		0	θ_o	$1-\theta_o$
9	1	0	0		0	$1-\theta_o$	θ_o
10	1	0	0		0	1	0
11	θ_o	0	$1-\theta_o$		0	θ_o	$1-\theta_o$
12	$1-\theta_o$	0	θ_o		0	$1-\theta_o$	θ_o
13	θ_o	0	$1-\theta_o$		$1-\theta_o$	θ_o	0
14	0	$1-\theta_o$	θ_o		θ_o	0	$1-\theta_o$
15	0	θ_o	$1-\theta_o$		0	0	1
16	0	$1-\theta_o$	θ_o		$1-\theta_o$	0	θ_o
17	$1-\theta_o$	θ_o	0		1	0	0
18	θ_o	$1-\theta_o$	0		θ_o	$1-\theta_o$	0

The remaining problem, for given θ_o, is a multiple regression problem. We are required to estimate H, M, L, b_2 and σ^2 for each specified value of θ_o. As we have indicated, this is a standard statistical problem and can be solved by employing standard statistical packages.

Table 3.9

Results of Experiments

Triple Design				Double Design				
WBC	RBC	Hb	MCV	WBC	RBC	Hb	MCV	time code
10^9/l	10^{12}/l	g/dl	fl	10^9/l	10^{12}/l	g/dl	fl	t
21.8	6.20	20.1	92	8.4	5.21	15.5	84	1
8.8	5.23	15.6	84	21.6	6.26	20.2	92	2
8.5	5.21	15.5	84	21.6	6.28	20.2	92	3
8.5	5.22	15.5	84	3.9	3.76	9.5	72	4
3.6	3.74	9.4	72	3.6	3.69	9.3	72	5
3.6	3.72	9.3	71	21.1	6.21	20.1	93	6
3.6	3.72	9.3	71	8.8	5.23	15.7	84	7
21.8	6.20	20.1	92	3.7	3.80	9.4	72	8
22.0	6.27	20.3	93	8.5	5.16	15.5	84	9
22.1	6.22	20.3	93	8.5	5.21	15.6	84	10
4.0	3.75	9.5	73	3.6	3.72	9.4	71	11
21.8	6.21	20.1	92	8.4	5.16	15.5	84	12
4.0	3.77	9.5	72	21.5	6.19	20.2	92	13
8.3	5.22	15.5	85	4.0	3.77	9.5	72	14
3.6	3.77	9.4	72	3.5	3.69	9.3	71	15
8.5	5.18	15.5	84	21.4	6.18	20.2	93	16
21.4	6.24	20.2	92	21.8	6.23	20.4	92	17
8.7	5.26	15.6	84	8.7	5.26	15.7	84	18

Table 3.9 gives the numerical results of the laboratory experiments which were carried out. Table 3.10 gives the Coulter Assay sheet values for 4C in Isoton II, which were used.

Table 3.10

Values from Coulter 4C assay sheet.

		WBC $10^9/l$	RBC $10^{12}/l$	Hb g/dl	MCV fl
Mean Values	H	22.0	6.07	20.0	94
	M	8.5	5.17	15.5	84
	L	3.4	3.65	9.1	71
Precision (St. dev.)	H	0.2	0.04	0.05	0.7
	M	0.1	0.02	0.05	0.6
	L	0.1	0.02	0.05	0.7

One of the assumptions implicit in employing the Triple or the Double procedures to estimate accuracy, precision etc., was that at each level the precision of the specific channel was identical, i.e. homogeneity of error variances. From the table given above we note that the values estimated by Coulter are not exactly the same for each level for the four clinical parameters of interest. However, the estimated differences are not sufficient to modify the analysis.

In the following tables, 3.11 and 3.12, we present the results of the analyses which were carried out on the data.

Having constructed point estimates of the various characteristics of each channel, we now require to construct

confidence intervals for the various estimates. The constructions given in the situation in which we could code the blood levels is not valid in this case. We require to use the properties of the type of estimator used. Statistical theory gives methods for obtaining the standard errors (approximate) of the various estimates. Hence we can construct approximate confidence intervals.

TABLE 3.11

Design: Triple
Clinical Parameter: WBC

Estimation	Point Estimate	Standard Error
High	22.03389	0.06611
Medium	8.48970	0.05383
Low	3.57641	0.06005
Carryover	0.02031703	0.00376
Drift	-0.00356	0.00290
Precision	0.12729355	

Design: Triple
Clinical Parameter: RBC

Estimation	Point Estimate	Standard Error
High	6.24295	0.00788
Medium	5.22268	0.00668
Low	3.72271	0.00820
Carryover	0.01663002	0.00360
Drift	0.00085	0.00037
Precision	0.01611080	

Design: Triple
Clinical Parameter: Hb

Estimation	Point Estimate	Standard Error
High	20.26164	0.01614
Medium	15.54052	0.01354
Low	9.31451	0.01660
Carryover	0.01495909	0.00169
Drift	0.00215	0.00074
Precision	0.03278053	

Design: Triple
Clinical Parameter: MCV

Estimation	Point Estimate	Standard Error
High	92.63642	0.26183
Medium	84.20711	0.22365
Low	71.48980	0.27187
Carryover	0.03054701	0.01402
Drift	0.01319	0.01216
Precision	0.53503181	

TABLE 3.12

Design: Double
Clinical Parameter: WBC

Estimation	Point Estimate	Standard Error
High	21.71430	0.05737
Medium	8.49388	0.04668
Low	3.55848	0.05241
Carryover	0.02050317	0.00333
Drift	0.00163	0.00252
Precision	0.11065434	

Design: Double
Clinical Parameter: RBC

Estimation	Point Estimate	Standard Error
High	6.25957	0.01264
Medium	5.20945	0.01078
Low	3.69932	0.01312
Carryover	0.02874357	0.00561
Drift	-0.00076	0.00059
Precision	0.02582443	

Design: Double
Clinical Parameter: Hb

Estimation	Point Estimate	Standard Error
High	20.30321	0.02000
Medium	15.59179	0.01682
Low	9.30500	0.02060
Carryover	0.01650331	0.00209
Drift	0.00315	0.00092
Precision	0.04066512	

Design: Double
Clinical Parameter: MCV

Estimation	Point Estimate	Standard Error
High	92.23890	0.22429
Medium	83.98875	0.18620
Low	71.77235	0.23307
Carryover	-0.00976233	0.01298
Drift	-0.00437	0.01040
Precision	0.45757151	

Comments on Results

We see from tables 3.11 and 3.12 that there are some discrepancies between the estimated values and the values reported in the Coulter Assay sheet, see table 3.10. However, if we compare the estimated values from the Double and Triple designs we see that any differences are small. One possible reason for these differences between estimated and Assay values is the fact that by the time the experiments were carried out, the High Abnormal 4C had passed the expiry data.

The estimates of precision are approximately the same or slightly less than the Coulter values. Drift appears to be approximately zero on all channels. The estimates of carryover, from all but MCV (Triple) suggest that carryover is under 3%. (In Section 5.8.9(b) we discuss the acceptable level of carryover, which is usually 3%). We note that the estimate of carryover for MCV (double) is negative. However, by using the estimate of standard error, the confidence interval is seen to include 0.0%, from which we can infer negligible carryover.

Finally we consider the covariance structure of the estimation procedure.

Covariance Structure.

The estimators we have given for θ, H, M, L and b_2 are equivalent to the maximum likelihood estimators obtained by maximising the logarithmic likelihood function, ℓ'. To construct this function we assume that the errors, $\varepsilon_t, t = 1,2,\ldots,18$ have independent identical Gaussian distributions. (See Wilks (B36), Kendall and Stuart (B23)). We maximise

$$\ell' = -\frac{1}{2}\ell n\ 2\pi - \frac{1}{2}\ell n\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{18}(y_t-(1-\theta)\Psi_t-\theta\Psi_{t-1}-b_2t')^2$$

with respect to θ, M, H, L and b_2.

The theory of maximum likelihood estimation shows that such estimators are asymptotically unbiased and have asymptotic Gaussian distributions. The variance and covariance of the estimators can be approximated by the following. We obtain the matrix A, of all second partial derivatives of ℓ' with respect to θ, H, M, L and b_2. The matrix A is symmetric. We give the upper triangular form.

$$A = \begin{bmatrix} \frac{\partial^2\ell'}{\partial b_2^2} & \frac{\partial^2\ell'}{\partial b_2\partial\theta} & \frac{\partial^2\ell'}{\partial b_2\partial H} & \frac{\partial^2\ell'}{\partial b_2\partial M} & \frac{\partial^2\ell^2}{\partial b_2 L} \\ & \frac{\partial^2\ell'}{\partial\theta^2} & \frac{\partial^2\ell'}{\partial\theta\partial H} & \frac{\partial^2\ell'}{\partial\theta\partial M} & \frac{\partial^2\ell'}{\partial\theta\partial L} \\ & & \frac{\partial^2\ell'}{\partial H^2} & \frac{\partial^2\ell'}{\partial H\partial M} & \frac{\partial^2\ell'}{\partial H\partial L} \\ & & & \frac{\partial^2\ell'}{\partial M^2} & \frac{\partial^2\ell'}{\partial M\partial L} \\ & & & & \frac{\partial^2\ell'}{\partial L^2} \end{bmatrix}$$

We take the expected value of the elements of matrix A, obtaining matrix B.

i.e. $B = E(A)$

Now the matrix B is inverted to obtain the matrix C. The diagonal elements of C give the approximate variances of the estimators, and the off-diagonal elements give the approximate covariances. (Note the matrices A, B and C are symmetric.)

For both designs the matrix B is

$$B = \begin{bmatrix} \frac{1938}{\sigma^2} & \frac{12(H-2M+L)}{\sigma^2} & \frac{12\theta}{\sigma^2} & \frac{-24\theta}{\sigma^2} & \frac{12\theta}{\sigma^2} \\ & \frac{8(H^2+M^2+L^2-HM-HL-ML)}{\sigma^2} R_H & & R_M & R_L \\ & & \frac{4(1-2\theta+2\theta^2)+2}{\sigma^2} & \frac{2\theta(1-\theta)}{\sigma^2} & \frac{2\theta(1-\theta)}{\sigma^2} \\ & & & \frac{4(1-2\theta+2\theta^2)+2}{\sigma^2} & \frac{2\theta(1-\theta)}{\sigma^2} \\ & & & & \frac{4(1-2\theta+2\theta^2)+2}{\sigma^2} \end{bmatrix}$$

where
$$R_H = \frac{y_H + 4(1-2\theta)(-2H+M+L) + 12b_2}{\sigma^2}$$
$$R_M = \frac{y_M + 4(1-2\theta)(H-2M+L) + 12b_2}{\sigma^2}$$
$$R_L = \frac{y_L + 4(1-2\theta)(H+M-2L) + 12b_2}{\sigma}$$

For the Double design.

$$y_H = y_2 + y_6 + y_{13} + y_{16} - y_4 - y_7 - y_{14} - y_{18}$$
$$y_M = y_7 + y_9 + y_{12} + y_{18} - y_2 - y_8 - y_{11} - y_{13}$$
$$y_L = y_4 + y_8 + y_{11} + y_{14} - y_6 - y_9 - y_{12} - y_{16}.$$

For the Triple Design.

$$y_H = y_1 + y_8 + y_{12} + y_{17} - y_2 - y_{11} - y_{13} - y_{18}$$

$$y_M = y_2 + y_{14} + y_{16} + y_{18} - y_1 - y_5 - y_{15} - y_{17}$$

$$y_L = y_5 + y_{11} + y_{13} + y_{15} - y_8 - y_{12} - y_{14} - y_{16}.$$

The inverse matrices are given in tables 3.13 - 3.14.

TABLE 3.13

Covariance - Drift, Carryover, H, M, L. Triple:WBC

0.00000840				
-0.00000077	0.00001417			
-0.00000856	0.00015040	0.00437057		
0.00000299	-0.00004186	-0.00046300	0.00289817	
0.00000557	-0.00010854	-0.00117063	0.00030176	0.00360580

Covariance - Drift, Carryover, H, M, L. Triple: RBC

0.00000013				
0.00000004	0.00001296			
0.00000004	0.00001519	0.00006203		
0.00000002	0.00000208	0.00000219	0.00004456	
-0.00000006	-0.00001727	-0.00002049	-0.00000302	0.00006724

Covariance - Drift, Carryover, H, M, L. Triple: Hb

0.00000055				
0.00000003	0.00000287			
0.00000012	0.00001494	0.00026050		
0.00000005	0.00000137	0.00000622	0.00018335	
-0.00000017	-0.00001631	-0.00008585	-0.00000870	0.00027542

Covariance - Drift, Carryover, H, M, L. Triple: MCV

0.00014787				
0.00000536	0.00019644			
0.00004306	0.00192547	0.06855502		
0.00002598	0.00025626	0.00200522	0.05001824	
-0.00006904	-0.00218173	-0.02188951	-0.00335274	0.07391297

TABLE 3.14

Covariance - Drift, Carryover, H, M, L. Double: WBC

0.00000635				
-0.00000058	0.00001110			
-0.00000629	0.00011509	0.00329088		
0.00000212	-0.00003017	-0.00032728	0.00217915	
0.00000418	-0.00008492	-0.00089518	0.00021655	0.00274705

Covariance - Drift, Carryover, H, M, L. Double: RBC

0.00000034				
0.00000009	0.00003143			
0.00000009	0.00003736	0.00015987		
0.00000006	0.00000482	0.00000463	0.00011621	
-0.00000014	-0.00004218	-0.00005124	-0.00000758	0.00017208

Covariance - Drift, Carryover, H, M, L. Double: Hb

0.00000085				
0.00000004	0.00000436			
0.00000020	0.00002272	0.00040012		
0.00000008	0.00000222	0.00001002	0.00028286	
-0.00000028	-0.00002494	-0.00013151	-0.00001425	0.00042438

Covariance - Drift, Carryover, H, M, L. Double: MCV

0.00010814				
0.00000410	0.00016855			
0.00004189	0.00163520	0.05030705		
0.00000060	0.00019524	0.00200610	0.03466963	
-0.00004249	-0.00183044	-0.01764570	-0.00200827	0.05432142

Comments on Computed Covariance Matrices.

The computed covariance matrices, see tables 3.13, 3.14, have, as expected small off diagonal elements. On average the largest correlation (approximately -0.3) occurs between the estimates of 'high' and 'low' levels.

We remark that these procedures are part of an evaluation programme. Basically this allows for simultaneous estimation of the parameters of interest. However, it will be obvious that if we wish only to calibrate the instrument, more efficient procedures are available. Further, the actual correlations depend on the 'distances' between levels. In section 3.10, we have shown the estimation to be orthogonal if the levels are equidistant. To achieve such a situation in practice is not particularly easy. However, given the abnormal blood specimens, it may be possible to arrange for adequate supplies of a 'normal' blood such that the levels are very nearly equidistant.

3.13 Construction of Designs : The General Situation.

3.13.1 Introduction.

The literature on statistical design contains only one specific solution to a similar problem to the one which we have discussed in the above sections. Daniel (1974) has given a design which has the properties which we require. The specific situation which Daniel investigated concerned the construction of designs for calibration of machines with carryover and drift. The specific machine he dealt with was the SMAC auto-analyser designed for use in Clinical Chemistry. This machine has some special properties. For example, after forty-eight measurements there is an automatic 'base-line' recalibration procedure. This single feature constrained the viable

designs, no matter the number of different 'levels' used, to have no more than forty eight measurements in total. Daniel was unable to give a general procedure for design construction - even in the three level case.

Some other work regarding 'systematic designs' has been reported by Cox (1951, 1952). For a review of some proposed systemmatic designs, with special reference to the removal of trend effects, proposed by Cox and Box, see Hill (1960). However, none of the designs are applicable in the particular situation which we have described. Perhaps the most relevent work has been by Williams (1952). He investigated designs which could be used for the analysis of 'treatments' in the presence of serial correlation. Due to the symmetry of this situation, if A and B are 'levels' the pairs AB and BA are equivalent. That is the sequence AB is the same as BA. Williams proposes designs in which either of equivalent pairs or both occur a specific number of times. The designs proposed cannot be used in the situations we have considered if we require the 'orthogonality' previously discussed.

In the following, we shall discuss the construction of experimental designs for the general situations. Both situations will be discussed - Codable levels and uncodable levels.

3.13.2 Codable Levels : The General Situation

We have investigated two specific sequences of three levels, and in the case of the Double design, we have indicated where, for other sequences, the analysis is modified. Now we consider how we might construct sequences satisfying the given constraints in the

general situation. This generalisation may be carried out in two ways. Firstly, we could extend the given designs to give 'larger' experiments. That is we could replicate the basic designs which have been given. This would lead to more measurements, and on average 'better' estimates of the parameters. Secondly, we could produce designs which use more than three levels. In order to investigate such constructions, we shall discuss the construction of designs for three levels more fully.

Let us consider the use of three levels i.e. three bloods specimens. The designs we have given are minimal sized designs under the constraints (i), (iia) and (iii). To obtain 'larger' experiments, we could replicate these basic designs. It will be seen that these new, larger designs also satisfy the constraints. On replication, the first level, which is used to obtain the 'missing pair', is not replicated. Replicating the Triple Design, we have

M : H M M M L L L ..... M H M | H M M M L L H ..... M H M .....

1st replicate | 2nd replicate

The incidence matrix, X, for such a design is easily constructed for both codable and uncodable situations.

In the codable case, the estimates in the reparameterised case is effected by using

$$\hat{\beta}' = (X'X)^{-1} X'Y$$

Replicating either the Double or Triple designs, say k times we have

$$(X'X)_k = \begin{bmatrix} 18k & 0 & 0 & 0 \\ 0 & 12k & 0 & 0 \\ 0 & 0 & 12k & 0 \\ 0 & 0 & 0 & 2\sum_{}^{9k} (2i-1)^2 \end{bmatrix}$$

This gives the inverse

$$(X'X)_k^{-1} = \frac{1}{k}\begin{bmatrix} 1/18 & 0 & 0 & 0 \\ 0 & 1/12 & 0 & 0 \\ 0 & 0 & 1/12 & 0 \\ 0 & 0 & 0 & \dfrac{1}{6(18k-1)(18k+1)} \end{bmatrix}$$

Thus, excepting the case of estimating drift, the variances of our estimates are changed by a factor of 1/k.

A mixture of these basic designs can also be used to produce larger designs. Suppose that we replicate the basic Triple design k_t times, and replicate the Double design k_d times, then in effect this is equivalent to either one of the basic designs replicated $k = k_d + k_t$ times. These experiments will have 18k + 1 measurements.

The earlier discussion has centred only on a single example of a Triple design and a single example of a Double Design. In the following section, alternative Double designs are considered. However, only one other Triple design has been found; this design being obtained by interchanging H and L in the Triple design given first in Section 3.8. (No other such transformations, nor reflections produce acceptable designs).

Other types of designs may exist, especially if we consider designs using more than three levels. For example, some levels may exhibit the 'triple' characteristic while others may exhibit the 'double' characteristic. We shall not investigate designs other than the Triple and Double designs already discussed and extensions to the Double Design. The general results given below concern only the Double designs and use the symmetric structures.

Consider designs using three levels, H, M, and L which are coded 1, 0 and -1 respectively. Further the designs are Double designs being symmetric and obviously balanced in time. The following argument shows that for such designs, a necessary condition for the design to be acceptable is that it begins with the 'middle' level, i.e. M.

Time is coded as before: -17, -15, -13,..., -1, 1, ..., 15, 17. Let h_i, m_i and l_i, i = 1,2,...,6. represent the time codes associated with the levels H, M and L respectively. That is, each of the h_i, m_i and l_i take one (unique) value from the sequence -17, -15,..., 15, 17. Thus the following relationships hold.

$$-h_1 = h_6 \qquad -m_1 = m_6 \qquad -l_1 = l_6$$
$$-h_2 = h_5 \qquad -m_2 = m_5 \qquad -l_2 = l_5$$
$$-h_3 = h_4 \qquad -m_3 = m_4 \qquad -l_3 = l_4$$

The term of immediate interest is $\Sigma \Psi_{t-1} t'$. By using the Double designs, which are symmetric and balanced in time and pairs, we only require to show that $\Sigma \Psi_{t-1} t' = 0$ to show orthogonality. (We have already discussed the benefits of orthogonality in estimation, see section 3.9.)

Assume that the sequence begins with M, the 'middle' level which has code 0. Let h_i', m_i' and l_i' represent the time codes of the preceding levels Ψ_{t-1}. Then we have:

$$h_i' = h_i - 2, \quad m_i' = m_i - 2 \text{ and } l_i' = l_i - 2 \quad i = 1,2,...,6.$$

From the previous relationships, we can write

$$-h_1' = h_6' - 4 \qquad\qquad -l_1' = l_6' - 4$$
$$-h_2' = h_5' - 4 \qquad \text{and} \qquad -l_2' = l_5' - 4$$
$$-h_3' = h_4' - 4 \qquad\qquad -l_3' = l_4' - 4$$

The cross-product time is:

$$\sum_{t=1}^{18} t' \Psi_{t-1} = (h_1'+h_2'+h_3'+h_4'+h_5'+h_6') - (l_1'+l_2'+l_3'+l_4'+l_5'+l_6')$$

This expression can be simplified to give:

$$\Sigma t'\Psi_{t-1} = (-h_6'+4-h_5'+4-h_4'+4+h_4'+h_5'+h_6') - (-1_6'+4-1_5'+4-1_4'+4+1_4'+1_5'+1_6')$$
$$= 12 - 12$$
$$= \underline{0}$$

Thus the cross-product term for such a design is zero when it starts with M.

Assume that the sequence does not begin with M. In fact it does not matter whether it begins with H or L, as the basic situation is symmetric in L and H. Thus we shall assume that it begins with H. With the same notation as before, in this case, $h_1' = -17$, but $-h_1' = h_6' - 4$ does not necessarily hold.

The cross-product term, for similar reasons as before, is

$$\Sigma t'\Psi_{t-1} = (-17-h_5'+4-h_4'+4+h_4'+h_5'+h_6') - (-1_6'+4-1_5'+4-1_4'+4+1_4'+1_5'+1_6')$$
$$= -12 + h_6'$$

This term is zero only if $h_6' = 21$. This can never be achieved, since the largest time code possible is 17. Hence, we cannot achieve orthogonality unless the sequence begins with the 'middle' level, i.e. M.

In the previous discussions, it was indicated that a time balance could be achieved easily for a sequence of eighteen levels if we merely repeat, in reverse order, the first nine levels. For example, we establish the first part of the sequence, up to the line, and then continue the sequence as shown.

M : M H H L L H M L M | M L M H L L H H M

Thus, if we start with the level M, and produce a sequence in which each pair of levels occur equally often, (excepting for MM which appears once at the beginning and will also appear at the axis of the

reflection), by reflection as above, the pairs will be balanced.

The construction of the sequence of the first nine levels plus the first unrecorded level can be effected in the following way.

Let us once again denote the three levels by H, M, and L. Construct a table such as

M	H	L		M	H	L
H	L	M	or	H	M	H
L	M	H		L	L	M
	(i).				(ii).	

We note that the first table, (i). is a 'cyclic' Latin square. There are two constraints on the possible tables.

(a). The level which appears in the top-left-hand corner must be M and at least one M must appear in the bottom row.

(b). Each column must have one and only one of the three levels.

The first row of these tables are used to denote the preceding level and the other rows, in order, are used to generate the sequence. The positions in which doubles will appear are marked thus *, with an asterisk. Thus, as examples for the above tables, we could have:

(i)			(ii)		
M*	H	L	M*	H	L
H*	L*	M	H*	M	H
L	M	H	L*	L	M

Let us consider the table (i). We start with a double M in the top-left-hand corner. The second row indicates that H (doubled) is next. The column headed H, has L in the second row (doubled), which is

therefore next in sequence. The column headed L, has M in the second row and M is next in the sequence. The column headed M has L in the third row - we have already 'used' the H* in the second row of this column. We proceed in a similar fashion obtaining the sequence:-

M : M H H L L M L H M

In example (ii) we proceed as above. Thus, starting with a double M, we have a double H next in sequence. However, we see that the column headed M has already had the entry in row two used. We proceed to the double L in the third row, but now we use the H in the second row of the column headed by L - before using the third row entry. The sequence is:

M : M H H M L L H L M

Both of these sequences are reflected as we have indicated previously, and we obtain the designs required.

We have shown that the design must start with M, and have also indicated that M must also appear in the final position. Then we must have tables which have at least one M in the final row. The only possible tables are the six which are listed below.

M	H	L
H	L	H
L	M	M

1.

M	H	L
H	L	M
L	M	H

2.

M	H	L
H	M	H
L	L	M

3.

M	H	L
L	L	H
H	M	M

4.

M	H	L
L	L	M
H	M	H

5.

M	H	L
L	M	H
H	L	M

6.

We can position the doubles for H and L in any of the positions in the second and third rows. However the double M must appear in the first position, i.e. M* must always appear in the top left-hand corner. It will be obvious why this must be. If it were not so, the cross-product $\Sigma t'\Psi_{t-1}$ would not have six of each level appearing as Ψ_{t-1}. All possible sequences are given in table 3.14.

Table 3.14

	M	:	M	H	H	L	L	H	M	L	M	M	L	M	H	L	L	H	H	M
	M	:	M	H	H	L	H	M	L	L	M	M	L	L	M	H	L	H	H	M
1.	M	:	M	H	L	L	H	H	M	L	M	M	L	M	H	H	L	L	H	M
	M	:	M	H	L	H	H	M	L	L	M	M	L	L	M	H	H	L	H	M
	M	:	M	H	H	L	L	M	L	H	M	M	H	L	M	L	L	H	H	M
	M	:	M	H	H	L	M	L	L	H	M	M	H	L	L	M	L	H	H	M
2.	M	:	M	H	L	L	M	L	H	H	M	M	H	H	L	M	L	L	H	M
	M	:	M	H	L	M	L	L	H	H	M	M	H	H	L	L	M	L	H	M
	M	:	M	H	H	M	L	L	H	L	M	M	L	H	L	L	M	H	H	M
	M	:	M	H	H	M	L	H	L	L	M	M	L	L	H	L	M	H	H	M
3.	M	:	M	H	M	L	H	H	L	L	M	M	L	L	H	H	L	M	H	M
	M	:	M	H	M	L	L	H	H	L	M	M	L	H	H	L	L	M	H	M
	M	:	M	L	L	H	H	L	M	H	M	M	H	M	L	H	H	L	L	M
	M	:	M	L	L	H	L	M	H	H	M	M	H	H	M	L	H	L	L	M
4.	M	:	M	L	H	H	L	L	M	H	M	M	H	M	L	L	H	H	L	M
	M	:	M	L	H	L	L	M	H	H	M	M	H	H	M	L	L	H	L	M
	M	:	M	L	L	M	H	H	L	H	M	M	H	L	H	H	M	L	L	M
	M	:	M	L	L	M	H	L	H	H	M	M	H	H	L	H	M	L	L	M
5.	M	:	M	L	M	H	H	L	L	H	M	M	H	L	L	H	H	M	L	M
	M	:	M	L	M	H	L	H	L	H	M	M	H	H	L	L	H	M	L	M
	M	:	M	L	L	H	H	M	H	L	M	M	L	H	M	H	H	L	L	M
	M	:	M	L	L	H	M	H	H	L	M	M	L	H	H	M	H	L	L	M
6.	M	:	M	L	H	H	M	H	L	L	M	M	L	L	H	M	H	H	L	M
	M	:	M	L	H	M	H	H	L	L	M	M	L	L	H	H	M	H	L	M

These sequences can be shown to produce schemes for orthogonal estimation. Further these designs are statistically equivalent.

3.13.3 Designs Using Three or More Levels.

Having discussed the use of three levels at some length, we now consider the use of more than three levels in such clinical experiments. It has been shown that when three levels are used, the minimal sized design under the constraints (i), (iia) and (iii) is eighteen plus one. This was shown by a specific argument for this case. We hypothesise that when b levels are used, the minimal sized design, under the same constraints, is of size $2b^2 + 1$. No proof has been found to support this hypothesis. However, we note that if a balance for pairs is to be achieved, the size of the design must be $hb^2 + 1$, where h is an integer. Thus to show that $2b^2 + 1$ is the minimal sized design, we need only show that the design of size $b^2 + 1$ is not viable and that a design of size $2b^2 + 1$ is viable. In fact if we insist that the design has the characteristics of the Double design, then it will be obvious that $2b^2 + 1$ is the minimum size for such a design. The following shows that for a design of the Double type, we can attain orthogonality only by using an odd number of levels, and further starting the sequence with the 'middle' level.

Let 1_{ij} denote the j^{th} time code of the i^{th} level. Let there be $b = 2k + 1$, levels, $k = 1,2,3,\ldots$. That is the number of levels used is odd. Under the hypothesis we have made concerning the minimal sized design, Double, there are $2b^2 + 1$ entries in the design sequence. At each level there are 2b replications.

The time codes are one (uniquely) of the sequence:

$$\pm 1, \quad \pm 3, \quad \pm 5, \quad \ldots\ldots, \quad \pm(2b^2 - 1)$$

The levels have codes:

$$0, \quad \pm 1, \quad \pm 2, \quad \pm 3, \quad \ldots\ldots, \quad \pm k$$

If the levels are put in ascending order, i.e. ascending order of the clinical parameter in which the interest is centred, the assignment of codes is in the order:

$$-k, \quad -(k-1), \quad \ldots\ldots, \quad -1, \quad 0, \quad 1, \quad 2, \quad \ldots\ldots, \quad (k-1), \quad k$$

Thus, the 'middle' level has code 0, and is denoted:

$$1_{k+1\ j} \qquad j=1,2,\ldots,\ 2b$$

Since the designs are Double and have the property of symmetry, the following relationships hold:

$$-1_{ij} = 1_{i\ 2b-j+1} \qquad i = 1,2,\ldots,k,\ k+2,\ldots,b \text{ and } j = 1,2,\ldots,2b$$

Let $1'_{ij}$ denote the time codes corresponding to the time-lagged levels Ψ_{t-1}. Then we have the relationships:

$$1'_{ij} = 1_{ij} - 2 \qquad i \neq k+1$$

$$-1'_{ij} = 1'_{i\ 2b-j+1} - 4$$

The cross-product term can be written in the following way.

$$\Sigma t'\Psi_{t-1} = -\sum_{i=1}^{k}(2i-1)\sum_{j=1}^{2b} 1'_{ij} + \sum_{i=1}^{k}(2i-1)\sum_{j=1}^{2b} 1'_{ij}$$

$$= -\sum_{i=1}^{k}(2i-1)\,b(4) + \sum_{i=1}^{k}(2i-1)\,b(4)$$

$$= 0.$$

Let some level other than the middle level begin the sequence. Let this be the p^{th} level where $p \neq k+1$. We consider the level to occur in the first half. Since the situation is symmetric this does not detract from the generality of the proof. The following is an outcome of starting with the p^{th} level.

$$1_{p\ 1} = -(2b^2 - 1)$$

Then the cross-product term is:

$$\Sigma t'\Psi_{t-1} = -\sum_{\substack{i=1\\ i\neq p}}^{k}(2i-1)\sum_{j=1}^{2b} 1'_{ij} + \sum_{\substack{i=1\\ i\neq p}}^{k}(2i-1)\sum_{j} 1'_{ij} - (2p-1)\sum_{j} 1'_{pj}$$

$$= -(2p-1)(-2b^2+1+1'_{p\ 2b}+4.2k) + 4(2p-1)b$$

$$= -(2p-1)4b + (2p-1)4b - (2p-1)\left[-4-2b^2+1+1'_{p\ 2b}\right]$$

$$= -(2p-1)\left[-2b^2+1-4+1'_{p\ 2b}\right]$$

For this term to be zero we require

$$1'_{p\ 2b} = (2b^2-1) + 4$$

But this cannot be achieved, since $2b^2-1$ is the largest code possible. Thus for the cross product term to be zero we must start the sequence with the 'middle' level.

The above proof that orthogonality can be achieved only if the sequence begins with the 'middle' level, relates only to an odd number of levels. By a similar type of proof, it can be shown that orthogonality cannot be achieved for an even number of levels.

The generalisation of the Triple design to more than three levels results in a genuine difficulty. By some rather elementary considerations, it is obvious that the longest allowable sequence of a level is three and we cannot consider quadruple designs or the like. However, as we have already stated, no general construction could be found for the three level case and no triple designs have been found for the larger designs. However, the Double design can be extended quite easily.

We have already given a procedure for constructing Double designs in the cases in which three levels are used. In a similar way, we can construct a procedure which will give orthogonal designs for any odd number of bloods (levels). Perhaps the easiest way to achieve this

is to use a table which is a 'cyclic' latin square. The only constraint on such a table is that the entry in the top-left-hand corner must be the 'middle' level. Let there be b = 2k+1, k = 1,2,3..., blood specimens. We denote by A_1 the 'middle' level, then the table is

$$
\begin{array}{ccccccccc}
A_1^* & A_2 & A_3 & A_4 & A_5 & A_6 & \cdots\cdots\cdots\cdots & A_{2k} & A_{2k+1} \\
A_2^* & A_3^* & A_4^* & A_5^* & A_6^* & \cdots\cdots & \cdots\cdots & A_{2k+1}^* & A_1 \\
A_3 & A_4 & A_5 & A_6 & A_7 & \cdots\cdots & \cdots\cdots & A_1 & A_2 \\
\vdots & & & & & & & \vdots & \vdots \\
A_{2k+1} & A_1 & A_2 & A_3 & A_4 & \cdots\cdots & \cdots\cdots & A_{2k-1} & A_{2k}
\end{array}
$$

where A_i represents the i^{th} level. We use Row 1 to denote the preceding treatment and the other rows, in order, to generate the sequence. By using the procedure previously described, we generate the sequence

$$A_1 : A_1\ A_2\ A_2\ A_3 \cdots\cdots A_{2k}\ A_{2k}\ A_{2k+1}\ A_{2k+1}\ A_1\ A_3\ A_5 \cdots\cdots A_{2k+1} A_2 \cdots$$

$$\cdots A_1 \quad \cdots\cdots A_5\ A_4\ A_3\ A_2\ A_1.$$

To obtain the full design, we use the 'mirror image' of this sequence.

In fact, orthogonal designs can be constructed from tables which are not cyclic latin squares. The constraints on these tables are that the middle level must be in the top-left-hand corner and must appear at least once in the bottom row. The columns of the table can be any arrangement of the levels $A_1, A_2, \ldots, A_{2k+1}$, which is possible in view of the other constraints.

For example, if five levels are used, A, B, C, D, and E with C the 'middle' level, two such tables and the corresponding sequences are:

C*	A	B	D	E		C*	E	A	D	B
A*	B*	D*	E*	C		B*	D*	D	B	C
B	D	E	C	A		D	A	B	A*	D
D	E	C	A	B		A	B	C	C	E*
E	C	A	B	D		E	C	E	E	A
		1.						2.		

1. C : C A A B B D D E E C B E A D C D A E B C E D B A C
C A B D E C B E A D C D A E B C E E D D B B A A C

2. C : C B B C D D B D A A D C A B E E D E A C E B A E C
C E A B E C A E D E E B A C D A A D B D D C B B C

It will be seen that both of these sequences satisfy the required design constraints.

3.13.4 Uncodable Levels: The General Situation

In the examples which we have given in 3.12 although the levels were not codable, the design used was one which we had constructed for a codable situation. However, we note that the term $\Sigma t' \Psi_{t-1}$ will in general not be zero. That is, orthogonality will not be achieved. The covariance structure in such cases, as exhibited by matrix C, will be such that the specific covariance terms are small. That is, the diagonal elements of C will be large in comparison to the off-diagonal elements. This means that in general we will achieve approximate orthogonality.

We conclude that for uncodable situations, the 'codable situation designs' should be employed, with the similar extension to the analysis which we have shown.

3.14 Conclusions.

The main investigations of this chapter, concerned the simultaneous estimation of accuracy, precision and carryover in the presence of drift (linear) for automatic measuring devices such as the Coulter-S. (Linearity can also be assessed if the levels can be coded.)

The current procedures for the estimation of these parameters is at best inefficient of time and resources, and at worse able to give misleading information. The designs which we have proposed allow for the simultaneous estimation of accuracy precision and carryover in the presence of linear drift. Two 'different' designs, which are statistically equivalent (when three levels are used), have been called the Triple design and the Double design.

Although these designs are statistically equivalent, in practice a technician may well rather use a Triple design, since with three levels replicated each three times he can immediately see a grossly out of control machine.

The Double designs, as we have shown, can easily be extended for cases in which more than three levels are used. No such extension has been found for the Triple design.

We propose that these designs should be used in investigations, e.g. after major overhauls of existing machines, or the evaluation of new machines.

The designs have a further possible use. They may be used as a part of an intra-laboratory quality control scheme. The practical experiments which we have reported, show that the use of these designs give estimates of indices which are usually being monitored. It is

appreciated that some computing facilities are necessary, but with limited facilities a single program can be used for all analyses.

An important question to answer, is, 'what size of design (experiment) should be used?' When we are considering the evaluation of a device, the advantage of using more than three levels, apart from obtaining better estimates, on average, of the indices, is that a better indication of (non) linearity is given. However, if, as we have hypothesised, the minimum design for b levels is $2b^2 + 1$, the size of design is dependent on double the square of the number of levels used. Thus it increases rapidly as the number of bloods indices are increased. As we have shown the possible sizes of symmetric Double designs are then 19, 51, 99, ... The basic problem is then to obtain sufficient blood specimens of sufficient quantity to satisfy the design requirements. If we assume as before that the Coulter uses approximately 1.4 ml of blood at every count, then we require b blood specimens of at least 2(1.4)b ml. One further point should be borne in mind. One of the basic assumptions which we have made in constructing the analysis, is that the drift is linear. In practice, this assumption may well only be justified with relatively small experiments.

As we have indicated, these designs may also be used as a component of quality control. In such circumstances, the only practical size of experiment is 18 + 1, i.e. the minimal sized design. (We have remarked that the Triple design of this size may be more practical.) The routine haematology laboratory will usually not have the resources to operate any larger design. Fortunately it appears that such a design size can provice reasonable estimates (point and interval) for the purpose of control.

4 Comparison of Machines/Methods

4.1 Introduction

In chapter 3, we discussed the evaluation of a continuous auto-analyser, the Coulter-S. However, many of the points made are pertinent when considering the evaluation of other clinical measuring devices. The discussion of the evaluation or assessment of the machine considered the device in 'isolation'. That is, although we considered assessing the machine by perhaps employing reference or standard materials, there was no direct comparison between the specific machine and other measuring devices which could have been used to measure the same clinical parameter(s).

In this chapter, we will frequently refer to the comparison between different machines, or the comparison between machines and other methods, i.e. routines, methodologies, which purport to measure the same clinical parameter(s). For a meaningful evaluation of any machine/method, we require to investigate not only its innate characteristics, see chapter 3, but also to compare its performance with existing (accepted) measuring processes.

The evaluation of a machine/method has two parts - see chapter 3. Both the physical and statistical evaluations should be carried out. Extending this, when comparing different measuring processes, comparisons should be made between firstly the physical characteristics of the two or more processes, e.g. through-put rates, through-put times, basic costs of operating, principles of measurement, standards which are available, etc. Secondly, we require to compare the actual results reported by the different processes.

In this chapter, we shall investigate methods of making statistical comparisons between different machines/methods. The physical comparison is basically qualitative, although there must be comparisons of the numerical quantities such as reagent consumptions, flow-rates, etc.

Usually, the measurement of any specific biological parameter can be effected by more than one methodology or procedure. Two problems arise. Firstly, when different methods/machines are employed, are we certain that exactly the same clinical parameter is measured? Secondly, even if the same parameter is measured, do the different methods/machines actually report the same, or equivalent results? Let us consider each of these questions in turn.

There are many examples in which quite different procedures are used to determine the 'same clinical parameter'. For example, the estimation of serum vitamin B_{12} can be carried out by a radio-isotopic method or by a microbiological method. As a second example, consider the estimation of haemoglobin in the peripheral blood. In this case, we must differentiate between total circulating haemoglobin, that is, all derivative forms of haemoglobin in a given aliquot, and the measurement of one or more of the specific haemoglobin derivatives. In this example, it is clear that the two alternative measurements do not reflect the same clinical parameter. This is not at all clear in the vitamin B_{12} example. Do the two different techniques actually measure the same clinical parameter? We shall call procedures which measure exactly the same clinical parameter equivalent measuring processes. All processes which do not have this property will be called non-equivalent processes. In the following sections we will only consider equivalent processes. In section 4.3 we return to consider the comparison of non-equivalent processes.

Let us now consider the second question which has been raised. If we investigate equivalent methods, how do we actually compare the processes? Basically two types of comparison could be made. Firstly, we could compare the results of the possible diagnosis for each patient based on the two processes. This of course entails some relationship between measurements and clinical status. This type of comparison is not without its problem as we show in 4.2. However, in the clinical sense this type of comparison is very important. The second type of comparison which could be made concerns the comparison of the actual numerical results which the different processes produce. This type of comparison is common, but as we show in section 4.3 has corresponding problems.

Before proceeding to these investigations, let us state the characteristics of the measuring processes which are admitted to the discussions.

(i). The processes are equivalent in the sense described above (see also 4.8).

(ii). They are single channelled i.e. only one result is reported. Multi-channelled machines can be handled by extending the discussion and results in very obvious ways.

(iii). The precision of the measuring process is high relative to the variability of the clinical parameter being measured. Generally the measuring error will be at least one order of magnitude smaller than the biological variability of the parameter being measured.

(iv)./

(iv). The errors associated with the method/machine have a Gaussian distribution and are independent of the specimen being measured, i.e. the errors are homoscedastic throughout the clinical range of measurement.

(v). The errors are independent of each other.

(vi). The bias, if any, is independent of the specimen being measured and is constant.

4.2 Methods and Indices of Comparison

4.2.1 Diagnostic Correlation

Unlike the mathematical approaches to comparison which we describe in section 4.3, we propose to correlate the clinical assessments of the various processes. The basic situation can be described as follows.

A standard method of classifying people as 'abnormal' and 'normal' exists. It is regarded as the absolute clinical method for such classification. The classification can have more than two categories. A new method is investigated, which theoretically should classify each person into the same clinical category as the standard method. The classification scale may however differ significantly in terms of the numerical values recorded for the clinical parameter measured by the methods. (In practice, it may well be that the same numbers i.e. the same scale of measurement is desirable. In such cases the clinical staff do not require to alter their 'normal ranges').

The comparison of two or more processes, does not require the numbers from the different processes to have exactly the same numerical values, only that they allow the same clinical diagnosis with the same, or better, reliability. In a sense, we wish to measure the

correlation of the diagnostic classifications.

Let us consider indices which may be constructed to reflect the correlation of classifications. Let us briefly indicate the ways in which we may collect the information on which to base the correlation indices. A comparison between two processes, each giving three categories is used in the explanation.

The table 4.1 corresponding to the situation is of the form:-

	New Method			
Standard Method	Abnormal Low	Normal	Abnormal High	
Abnormal Low	n_{11}	n_{12}	n_{13}	$n_{1.}$
Normal	n_{21}	n_{22}	n_{23}	$n_{2.}$
Abnormal High	n_{31}	n_{32}	n_{33}	$n_{3.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

TABLE 4.1

The total number of patients who have been classified by both methods is n. The following points are noted.

(i). The marginals, $n_{.1}$, $n_{.2}$ and $n_{.3}$ will always be random variables, since we can have no control over the classification of the new method.

(ii). If the n patients have been chosen at random, the marginals $n_{1.}$, $n_{2.}$ and $n_{3.}$ will be random variables reflecting the proportion of each classification in the underlying population.

(iii). A different type of sampling could be used. We could sample from each of the sub-populations - Abnormal High, Normal and Abnormal Low as specified by the standard method. In this case the marginals $n_{1.}$, $n_{2.}$ and $n_{3.}$ and n are fixed.

It is important to note these points if the distribution of functions of these data are required.

In situations where measurements are made to give preliminary information to the clinican, or in screening situations the number of categories for each method will usually be two or three. However, this approach can be used for more than three categories. In the following discussion we will mostly consider two or three categories.

4.2.2 Traditional Measures of Association

We give a brief review of some classical measures of association.

2X2 Tables.

To begin the discussion we consider two categories for each method. The table in this case has the following form, see table 4.2

	New method	
Standard Method	Positive (+)	Negative (-)
Diseased	a	c
Not Diseased	b	d

TABLE 4.2

Yule (see Kendall and Stuart (B)) has proposed the measures:-

$$Q = \frac{ad - bc}{ad + bc}$$

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} - \sqrt{bc}} \qquad \left(\text{note } Q = \frac{2Y}{1 + Y^2}\right)$$

Another measure, related to the χ^2 statistic discussed later is:-

$$V = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Youden (1950) proposed a measure which was based on the sensitivity and specificity. These terms are defined in the following way:-

$$\text{Sensitivity} = \frac{a}{a + c}\,\% \qquad \text{Specificity} = \frac{d}{d + b}\,\%$$

From these definitions, sensitivity is a measure of the ability of the new method to classify diseased patients as diseased, whereas, specificity is the ability of the new method to classify 'healthy' as healthy.

Youden's measure is:-

$$J = \tfrac{1}{2}\left(\frac{a - c}{a + c} + \frac{d - b}{d + b}\right) = \text{Sensitivity} + \text{Specificity} - 1$$

Now, although this measure is not a traditional measure of association, in the particular situation of comparing the diagnostic classifications it can be thought to have this interpretation. The measure J is the average of the success of the new method on the diseased and healthy groups. If the new test had no discrimination for diseased people then a - c would be zero. In fact this measure supposes that false positives, 'b', and false negatives, 'c', are equally undesirable. We comment on this later. For the present we describe indices which may be used with any general number of classifications.

4.2.3 General rXc Tables

In the introductory remarks and in the last section, we have discussed square tables ie rXr with r = 2 and r = 3. Square tables occur when the methods being compared can classify the 'measured patients' into the same number of categories. There may be situations in which one of the methods could classify into more categories than the other. For example, one may classify into r categories while the other classifies into c categories. In such a situation we would have a rXc table. In the following discussion we shall consider the general rXc table. (Note the square tables rXr are just special cases).

Probably the most common measure in the general situation, rXc,

is the familiar quantity denoted χ^2. This quantity is:-

$$\chi^2 = \sum_i^r \sum_j^c \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}}$$

which is the summation of $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ in each cell, under the assumption of independence.

In attempts to normalise this possibly infinite measure, the following modifications have been suggested.

$$\phi^2 = \frac{\chi^2}{n}$$

A variation by Pearson (see Kendall and Stuart(B23)) is

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}}$$

has been called the coefficient of contingency. Tschuprow (see Kendall and Stuart(B23)) has suggested the use of

$$T = \sqrt{\frac{\chi^2/n}{(r-1)(c-1)}}$$

especially in the case of rectangular tables. However, a 'better' measure suggested by Cramer is

$$C_c = \sqrt{\frac{\chi^2/n}{\min(r-1, c-1)}}$$

A measure, t_b, has been proposed by Stuart (1953) which is based on ranking procedures. Essentially t_b is a modified rank correlation index. The modification is required as we have a situation of many ties. For computational procedures see Kendall and Stuart, (B23) and Stuart (1953).

The review of some proposed measures given above would lead us to suspect that there is no shortage of 'measures'. If so many measures

are available, what characteristics would we wish a measure of association in this situation to possess?

Let us consider in more detail the properties which indices of association possess. In particular, we shall give two alternative ways in which indices can be scaled.

(i). The indices can be scaled to lie between 0 and 1 inclusive. An index has the value 1 if exact correlation exists between diagnostic results. The index has a value 0 if there is a random effect ie independence.

(ii). The indices can be scaled to lie between -1 and 1 inclusive. An index has the value 1 if there is complete agreement between methods. If there is complete disagreement between methods the index has value -1. When classification is random the measure has value 0.

Now let us consider the proposed indices, which have been given above. Our investigation centres on the effectiveness of indices to reflect correlation between methods.

The measures Q and Y are unity if <u>any</u> cell is zero. For example consider the following data presented in table 4.3.

New Method / Old method	Abnormal	Normal	
abnormal	20	10	30
normal	0	40	40
	20	50	

TABLE 4.3

Then we have Q = Y = 1

This type of behaviour is unreasonable in our situation. The index 'infers' complete association, which is not the inference we would

presumably wish to make.

The index, V, has perhaps a more acceptable behaviour in such a case. In the example above $V = (20 \times 40 - 10 \times 0)/\sqrt{20 \times 50 \times 40 \times 30}$ = 0.730, which infers a correlation but not an exact one. As we shall see, the index V is perhaps of more use than either Q or Y in this case. V will take a value -1 if a and d are simultaneously zero, and will attain 1 only if b and c are simultaneously zero.

The index, J proposed by Youden is scaled to lie in the range 0 to 1. J is 0 if the same proportion is classified positive for both diseased and not diseased groups. J takes the value 1 if there is an absence of false positive and false negative. The value of J is independent of the relative and absolute sizes of the diseased and healthy groups. All new methods giving the same value of J against a particular standard will make the same number (total) of misclassifications. Further if we assume that the number in each group is larger than 20 we can obtain an approximate expression for the standard error of J:-

$$SE(J) = \sqrt{\frac{a}{(a+c)^3} + \frac{d}{(d+b)^3}}$$

We note that the usual χ^2 index is not scaled in the sense of either (i) or (ii). Further any attempt, even with large cell values, which should not be expected in the off-diagonals, to give a χ^2 distribution property is incorrect. We do not consider that χ^2 is a reasonable index in this situation.

The use of the index t_b is probably better justified when there are a large number of categories. It is scaled to lie between -1 and 1. In the two or three category case, the computation is more difficult than the other measures, and is not really justified in terms of the amount of information it yields. In section 4.2.5 we contrast and compare the numerical values of some of these indices.

At this stage it is worth while to consider exactly what we understand by association. To do this let us consider the following examples given by Goodman and Kruskal (1954). The tables presented are 3X3 and have the same structure as that presented in section 4.2.1.

Example 1

n_{11}	0	0
0	n_{22}	0
0	0	n_{33}

Example 2

0	0	n_{13}
0	n_{22}	0
n_{31}	0	0

The pattern given in example 1 is that of complete association. The pattern shown in example 2 is one of complete disassociation, i.e. an inversion of the clinical groups.

Example 3

n_{11}	0	0
0	n_{22}	0
0	n_{32}	0

Example 4

n_{11}	0	0
n_{21}	n_{22}	0
0	n_{32}	n_{33}

In some circumstances the pattern presented in 3 may be considered as exhibiting association, but we shall consider neither 3 nor 4 to represent complete association.

4.2.4 Reliability Models

In the previous sections, we have reviewed and discussed some of the classical indices used for measuring association in contingency tables. As we have indicated some of these indices have properties which are not appropriate in the situation under discussion. Further, we have seen that in certain circumstances different definitions of association could be admitted. However, in the majority of situations in clinical comparisons, we find that interest centres on a particular tabular structure.

We are essentially interested in what Goodman and Kruskal (1954) have called reliability models. Two different methods of assignment (classification) of the same polychotomies are investigated. The indices which are constructed attempt to reflect to what degree the methods agree. A similar situation has been investigated in psychological contexts. However, the two methods of classification are subjective and neither can be considered as being 'true' necessarily.

In our particular situation the standard method is considered to give the correct classification, whereas the new method may not. (In some cases the standard method may be known to misclassify certain patients, but in these investigations we consider that such pathological patients are not included).

Many of the classical indices are invariant to permutations of rows and columns. This particular property is not desirable in our particular situation. To this end we consider two further indices which have been proposed by Goodman and Kruskal. These authors have suggested that the only reasonable indices are based on either the probability of correct classification or the probability of agreement to within one or two neighbouring categories. The indices are:-

I. $\sum_i \frac{n_{ii}}{n}$

II. $\sum_{|i-j|<k} \frac{n_{ij}}{n}$ where k is usually a small integer. When k = 1 we have I.

4.2.5 Numerical Comparisons of Indices.

We present below the values of some of the indices which we have discussed for five particular tables.

Ex. 1	30	0	0	30	Ex. 2	20	10	0	20
	0	60	0	60		10	50	0	60
	0	0	10	10		0	0	10	10

Ex. 3	15	15	0	30	Ex. 4	0	30	0	30
	15	45	0	60		0	60	0	60
	0	0	10	10		0	0	10	10

Ex. 5	25	5	0	30
	5	50	5	60
	0	5	5	10

Table 4.4 gives the numerical values of some of the measures which we have discussed in relation to rXr tables.

Example	C	T	C_c	Goodman/ Kruskal I	t_b
1	0.816	0.707	1.000	1.000	1.000
2	0.759	0.583	0.820	0.800	0.815
3	0.716	0.512	0.725	0.700	0.583
4	0.775	0.612	0.866	0.700	0.548
5	0.581	0.357	0.505	0.800	0.815

TABLE 4.4

Firstly, we should note the inability of C and T to reflect complete association, i.e. example 1. Secondly, we note that the measures of Goodman and Kruskal, C_c and t_b seem to have similar values, except in example 5. The easiest of these measures to calculate is quite definitely the Goodman and Kruskal index. Hence this simple index which measures the degree of correct assignment of the new method as compared with the standard method appears to be useful and easy to calculate. Let us now consider some 2X2 tables and consider indices specifically of use in these cases and also C_c and Goodman and Kruskal's index.

Example 6.	40	10	Example 7	20	10
	10	40		10	60
Example 8	50	0	Example 9	40	20
	0	50		0	40
Example 10	40	0	Example 11	35	10
	20	40		15	40

The numerical results for the indices are given in table 4.5.

TABLE 4.5

Example	J	V	C_c	Goodman/Kruskal I
6	0.600	0.600	0.600	0.800
7	0.524	0.524	0.524	0.800
8	1.000	1.000	1.000	1.000
9	0.667	0.667	0.667	0.800
10	0.667	0.667	0.667	0.800
11	0.505	0.503	0.447	0.750

As expected in the simple 2X2 tables the numerical values for the different indices tend to be very similar. In fact equal in some circumstances. In this situation, because of the practical interest in specificity and sensitivity the J index, as well as appearing to give reasonable reflection of the situation, has some clinical basis. Goodman and Kruskal's index gives only a very rough impression of the structure of the table.

4.2.6 Loss Function Approach.

In this section, we discuss a loss function approach to the comparison of different clinical classifying methods. We deal only with 3x3 tables. That is each of the two methods being compared, 'old' and 'new', can classify into three categories; 'abnormal low', 'normal' and 'abnormal high'. The notation used was first introduced in section 4.2.1.

Let us consider that l_{ij} is the 'loss' associated with classifying a person, known to be a member of the i^{th} class as measured by the 'old' method, in the j^{th} class by the 'new' method. We can estimate the probability that such a loss is incurred this being $n_{ij}/n_{i.}$. Hence the Average Expected Loss is $\sum_{ij} \frac{n_{ij}}{n} l_{ij}$.

Construction of the Loss Function, l_{ij}

The loss function imposed in any situation, will almost certainly be purely subjective. For this reason, many workers, including Goodman and Kruskal, although attracted at first by this approach, reluctantly reject it, mainly due to the difficulty in constructing a meaningful and consistent loss function, i.e. assigning values to the l_{ij}.

Let us investigate the situation more fully. The loss, l_{ij}, could be likened to the actual cost involved in misclassifying a patient, when $i \neq j$, and to correctly classifying a patient when $i = j$. Commonly the cost of correct classification l_{ii}, $i = 1, 2, 3$, will be zero. (We note that it still 'costs' to classify correctly, but this cost will be common to all classifications and can be set equal to zero.) The cost, although most easily thought of in terms of money, does not necessarily have to be specified in such a manner. It would reflect the cost of either treating or further investigating healthy or 'normal' persons, or not treating 'diseased' persons. This last type of misclassification creates problems concerning the cost or loss in prolonged disease or even death.

If we consider the 'low abnormals', we note that the losses l_{12} and l_{21} will not in general be equal; l_{12} is the loss associated with a 'low' false negative and l_{21} is the loss associated

with a 'low' false positive. In general the other losses such as these will not be equal either. The loss associated with a 'high' false positive is l_{23}, l_{32} is the loss associated with a 'high' false negative. Thus in most cases the costs of these different misclassifications will be different, i.e.

$$l_{21} \neq l_{12},\ l_{23} \neq l_{32},\ l_{32} \neq l_{12},\ l_{23} \neq l_{21}$$

In general the loss l_{ij}, where $|i - j| > 1$, will be enormous. This is the loss involved by classifying a person into a category not even neighbouring the 'correct' category. This will, in general, be a catastrophic misdiagnosis.

Even after constructing a loss function for a specific situation, i.e. comparison, we note that such a loss structure is only relevant for that specific situation. Thus the loss structure requires to be defined for each investigation of different classifications.

4.2.7 The Use of Loss Approach

This approach would appear reasonable if we had one of the following situations

(i). We are comparing a number of 'new' methods with a standard method. (This includes assessing 'quick and dirty' methods with standard methods, e.g., quick screening tests.

(ii). We are comparing a standard method and a new proposed method. There is also a reference method, perhaps too long or costly to be implemented in the routine laboratory.

As we are in a situation of making comparisons, it would appear reasonable to use the average loss, defined above, as an index for comparison. This is valid since the same loss structure is implemented.

Let $l_{ij} \geq 0 \; \forall \; i,j$, then obviously we have

$$0 \leq \text{Average Loss} \leq \frac{1}{n} \sum_{ij} n_{ij} \max_{j}(l_{ij})$$

The upper limit is attained when all persons are classified in the most 'expensive' way. This maxiumum will, in practice, never be attained. Such a method would have been rejected out of hand at a much earlier assessment stage.

The average loss obtained from each classification can now be compared. Such an index cannot be used as an absolute measure of acceptability. No consistent method of assigning losses is possible. We note two points of practical importance. Firstly, such an index is specific to a set of comparisons. Secondly, as in most comparative experiments, it is common practice to measure the same materials in each method being investigated. Thus by comparing each 'new' method against the 'accepted' method, we obtain the same 'old/accepted' method marginals.

Loss

We have considered the loss associated with misclassification. In other fields a great deal has been written about loss and optimal decisions. In such work in place of 'loss', utilities are used. Essentially a utility is a measure of how desirable a consequence is. Further the utilities are in fact probabilities. There is obviously a similarity between the use of utility and 'loss'. We note, however, that utilities are measured on a probability scale. Utility and loss are equivalent when utility is linear (see Lindley (B25)). Losses are in fact differences in utilities times a factor.

Although this approach, using a loss function, has not been followed here, it deserves to be investigated further. However, we

have already mentioned that it would be inconceivable to reject or accept a method/machine in the light of one index. We discuss in the following sections more comprehensive approaches to the comparison of methods/machines. We shall deal with the comparison of equivalent methods/machines.

4.3 Correlation

Let us consider the situation in which we compare two methods/machines, measuring a single clinical parameter. In this situation many authors have employed classical simple linear regression and correlation analysis. We shall investigate this approach and show that it is not the most appropriate, and further that in most situations it will produce misleading results.

Let us denote by the random variables Y and X the parameter determinations as measured by method 1 and method 2 respectively. Commonly method 2 will be the 'old' or standard method, and method 1 will be the 'new' method. We take n specimens, which hopefully cover the complete clinical range of interest. We measure each specimen using both method 1 and method 2. The resulting data is:-

$$\{(x_1,y_1), (x_2,y_2), (x_3, y_3), \ldots\ldots, (x_n,y_n)\}$$

As a first step, a 'scatter diagram' should be constructed, which will often have the following appearance, see figure 4.1

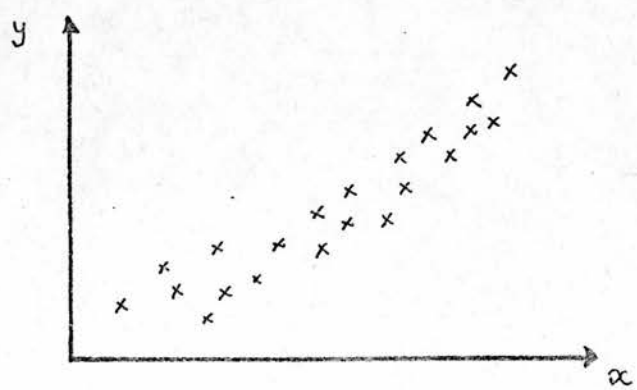


FIGURE 4.1

If an exact relationship existed, and no 'errors' were present, we would have all points lying on a line, not necessarily straight. However, due to experimental error (error in measurement) or to the relationship not being correct, the points are scattered as in the above diagram. We have defined previously what we mean by the error in measurement.

If such a 'scatter diagram' has been found, correlation of the measurements is obvious. We must remember in the following discussion that there is a difference between clinical and statistical correlation. We have discussed the concept of diagnostic correlation earlier, (see 4.2.1).

A common index to summarise the association is the product-moment correlation co-efficient i.e.

$$r_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{(n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2)}} = \frac{S_{xy}}{S_x S_y}$$

where S_{xy} is the unbiased estimate of covariance and S_x^2 and S_y^2 are that unbiased estimate of population variances.

This index has the following properties:-

(i). $-1 \leq r_{xy} \leq 1$

(ii). If $r_{xy} \approx 1$ a strong positive correlation exists in the data.

(iii). The correlation 'measured' by this index is purely linear.

To illustrate (iii), consider 'sampling' the data in table 4.5, which satisfies the exact relationship, $Y = X^2$.

Y	4	1	0	1	4
X	-2	-1	0	1	2

TABLE 4.5

The pictorial representation is given in figure 4.2.

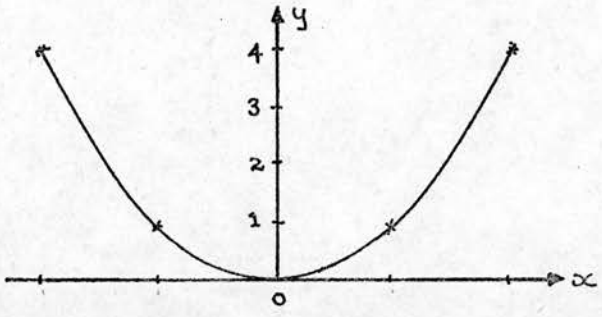


FIGURE 4.2

The product moment correlation coefficient, r_{xy} is zero in this case. We note, however, that an exact relationship exists between Y and X, i.e. Y and X are directly related. Thus although we can argue that when $r_{xy} \approx 1$ a relationship holds within the data, the fact is that when r_{xy} is not 1, even if it is zero, this does not necessarily imply that no relationship exists. The index, r_{xy} reflects the linear relationship within the data. The example given here exhibits a quadratic relationship. Kendall and Stuart, (B 23), remark that

> 'r_{xy} expresses the interdependence of X and Y, it does not measure it. .... the problem of joint variation is too complex to be comprehended in a single coefficient.'

Further the value expressed by r_{xy} depends on the sampling method as well as the strength of the association. To see this, consider the example given above, i.e. data satisfying the relationship $Y = X^2$. Suppose we sample the data given in table 4.6

Y	0	1	4	9	16
X	0	1	2	3	4

TABLE 4.6

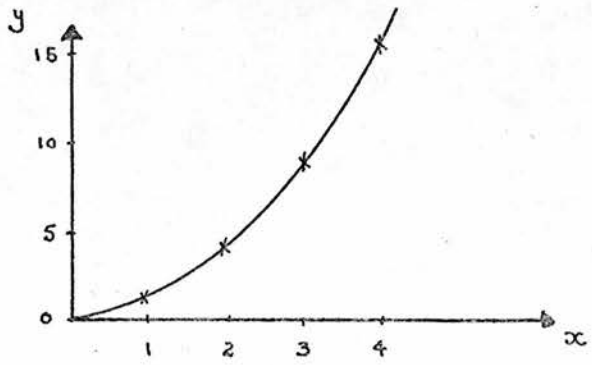


FIGURE 4.3

The graphical representation is given in figure 4.3. The product moment correlation coefficient, r_{xy} in this case is 0.9589. This is completely different from the previous value of 0.

As a practical example of the problems of interpretation of the product moment correlation coefficient and the effect of 'sampling', we consider the discussion given by Reed (1972). Reed criticises the use to which Berenson, Srinivasan, Lopez, Radhakrishnamurthy, Pargaonkar and Deupree (1972) put the correlation coefficient. The situation in which they used the correlation coefficient was in the comparison of different methods for measuring serum β and pre - β - lipoprotiens. Reed shows that the correlations exhibited could be misleading. He emphasises the point that he is not doubting that a relationship between methods exists. To show more clearly the basic problem he gives another example dealing with the comparison of two methods for determining phenylalanine in mg per cent. Twenty six patient specimens were measured by both methods, giving values for both methods in the range 1.8 to 3.3. Two control specimens, one at a high and one at a low level were also measured. The high produced values of 8.0 and 8.5, while the low gave values very close to the other 26 'normal' measurements. Reed calculated the correlation coefficient for the total data set (28 pairs) and obtained a value 0.982. However, when the correlation coefficient was calculated for the 27 data points - all data minus the high control, the value was 0.681. Reed concludes that by choosing certain control specimens, we could increase r_{xy} spectacularly. Again, no doubt is cast on a possible linear relationship between the methods, only that inference using r_{xy} is only justified within the statistical framework.

Although we have hypothesised a relationship to exist between methods, this does not necessarily constrain the relationship to be linear. Often the relationship will not be linear. In the particular clinical situation with which we deal, a linear relationship will most probably exist for at least some range of clinical values. Usually the measuring techniques will be devised in such a way, that the linear range will coincide with the 'normal range' of the clinical parameter. However, the comparison of methods with respect to pathological or abnormal specimens is of importance. If, as is common for practical reasons, the specimens used to compare methods come predominantly from the 'normal range', any non-linearity may well be masked by simply looking at r_{xy}. In these situations, we are interested in distinguishing situations (b), (c), (d) and (e) from (a) see figure 4.4

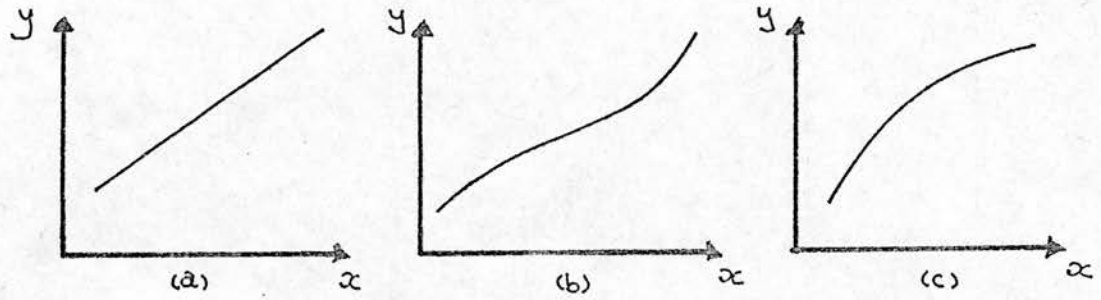
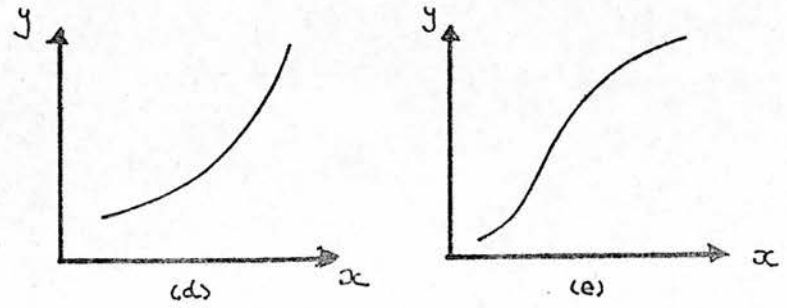


FIGURE 4.4



The situation represented by (a) is the ideal situation. The straight line does not always have to be the identity line, y = x. A recalibration or some correction factors can be used to give equivalent results on

both methods.

We note that different measuring methods may be linear over different clinical ranges. In practice, it would be beneficial to have a method which was linear over the complete clinical range - pathological and normal ranges. Non-linearity produces variable ability to detect specific clinical values. If a method is linear with respect to the true value of the clinical parameter being measured, then a change of one unit any place in the range reflects a change of exactly the same amount no matter where it is in the clinical range. If the relationship is not linear this is not true.

There is another problem, which is potentially troublesome. When we described the major characteristics of the measuring processes (section 3.5), we noted (iv) that throughout the clinical range, we had constant precision. This is an assumption which is of fundamental importance in the standard statistical analyses. Practically, this is an assumption which will not always hold. However, by employing a variance stabilising transform we can usually produce a data set which has this property and so carry out a valid statistical analysis. Consequently, we shall not consider this problem further here.

We conclude that at best the product moment correlation coefficient reflects the possible association. However, as we have shown, it is affected by sampling. In any situation, complete reliance on the correlation coefficient, is to be strongly discouraged. The common practice of calculating this coefficient and subsequently carrying out a hypothesis test to determine if the population correlation coefficient is different from 0 should be abandoned, as it makes no sense to investigate the null hypothesis of no association between methods.

As we have noted, any attempt to measure such a complex concept such as correlation by one index, is almost certain to fail. In the following sections, we discuss approaches which have been used.

4.4 Regression

Most investigators, faced with the problems raised above, try to approach the analysis in a different way, prefering to identify the 'relationship' between the measurement methods. In fact, they do not usually succeed in doing this as we show below.

Consider the use of the classical simple linear regression model. The model is:-

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1, 2, 3, \ldots, n$$

where ε_i is a Gaussian random variable with mean 0 and variance σ^2, and these errors are independent of each other and of the value x_i. The model identifies the error ε_i with the random variable Y; it is the part not explained by $\alpha + \beta x_i$. The conditional distribution of y_i given x_i is Gaussian with mean $\alpha + \beta x_i$ and variance σ^2. The problem is that such an approach would imply that we knew the exact value of X and that it was x_i. This, however, is not the case in practice. The observations x_i and y_i are measurements made on the unknown values X_i and Y_i. The following reasons have been put forward to justify the use of this particular model.

(i). The 'old' method, X, has been defined as the standard and by definition produces the 'correct' value, i.e. is without error.

(ii). The error in one of the measurements, X, is small in comparison with the corresponding error in the other, Y.

(iii). The error in X is small in comparison with the biological variability of X.

The following counter arguments must be recognised. No method can produce exactly repeatable results, hence the idea of an error-free measurement is just not tenable. Further, we cannot, without further information, suppose that one method has greater precision than the other. We shall show later, that the use of this type of analysis will result in wrong and possibly misleading estimates of the statistical parameters of the model, and hence to wrong and possibly misleading estimates of the clinical parameters. If the assumption under which we introduced the model hold, then it can be shown that the estimates of α and β are given by a and b respectively, where

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

If the situation is one in which there is an error in x_i, but the same analysis is carried out, the regression coefficient b is an estimate of

$$\frac{\beta}{1+\lambda} \quad \text{where } \lambda = \frac{\text{Variance of error in X}}{\text{Variance of X}}$$

That is b is always an under estimate of the 'slope' of the line. We note that even when the condition is that the error variance is at least one order of magnitude less than the biological variability is satisfied, this does not imply that a good estimate is obtained. In fact, if for example the true value of β was 1.05 and the error variance was one tenth of the biological variability, then the quantity we actually estimate has an expected value of 0.9545. That is in this situation the regression coefficient, β, is biased towards zero.

Further, it can be shown, see Bloch (1978), that this also tends to lower the probability of rejecting the 'null hypothesis' that the regression coefficient is zero.

The error identified by the simple linear regression model is really made up from two sources. The first is the error in measurement of the dependent variable. The second is the error in the model which has been assumed. This point will be developed in section 4.5.1.

There is a special situation when errors exist in both variables and regression can be justifiably used. This has been called the Berkson Model. We describe the basic situation for its use in the following.

Suppose that the 'true values' of x_i and y_i are X_i and Y_i respectively. Then due to measurement error in both observed variables, we have,

$$x_i = X_i + \delta_i \quad \text{and} \quad y_i = Y_i + \varepsilon_i$$

The usual regression model specifies that y_i be regressed on the errorless X_i. In fact the regression line is written as

$$y_i = \alpha + \beta X_i + \varepsilon_i \qquad i = 1, 2, 3, \ldots, n$$

What we really experience is

$$y_1 = \alpha + \beta(x_i - \delta_i) + \varepsilon_i$$

If we let $\phi_i = \varepsilon_i - \beta\delta_i$ we can write y_i in the following way:-

$$y_i = \alpha + \beta x_i + \phi_i$$

We note that this is similar to the simple regression case. However, in fact it is not the same. A correlation exists between x_i and 'the error term' ϕ_i. This invalidates the simple regression analysis.

Berkson showed, and his work is reviewed by Mandansky (1959) and Kendall and Stuart (B 23), that in certain circumstances, the situation

may reduce to the simple regression model. This would happen if the variable x_i were 'controlled'. This can arise in experimental situations. For example, consider a situation in which a machine is set to predetermined values during the experiment. Due to error of the machine, the actual 'true' values of the setting are different from the 'reported' values. It follows that in this special case, because the x_i are fixed, there is no correlation between x_i and ϕ_i. Thus in this special case the usual slope estimate is an unbiased estimated of β. We shall not pursue this further, as in clinical investigations, such as we are considering the variables cannot usually be controlled in this way.

The least squares procedure in 'regression analysis' identifies error in a single variate.

Under the usual regression model,

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1, 2, 3, \ldots, n$$

the error ε_i is associated with the random variable y_i. The estimating procedure for the parameters α, β and σ^2 is generally least squares. That is the values of α and β which minimise the error sum of squares,

$$\Sigma\varepsilon_i^2 = \Sigma(y_i - \alpha - \beta x_i)^2$$

This procedure is justified since we are minimising the error ε_i which is the error in y_i. If both variables have error, the analysis is unjustified, since we are minimising the error sum of squares with respect to a particular variate.

Instead of minimising in directions parallel to the y-axis or x-axis, we could obtain the 'orthogonal regression'. That is, minimise the perpendicular distances from the data points to the line. Figure 4.5 shows pictorially the distances d_i which are minimised to

give the estimates of α and β.

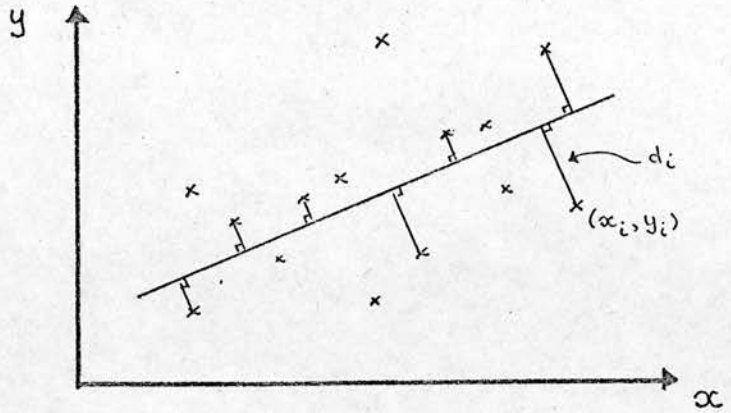


FIGURE 4.5

Denote by δ_i the measurement errors in x_i, i = 1, 2, ..., n. In the following sections we shall discuss methods for approaching the analyses of such data, which are dependent on the form of information concerning the error variances.

For present, let $\phi = V(\varepsilon)/V(\delta)$, i.e. the ratio of the variance of errors in y to the variance of errors in x, and assume that ϕ is known. Then a method of obtaining estimates of α and β is to minimise the weighted sum of squares

$$S_w = \Sigma\{(y_i - \hat{y}_i)^2 + \phi(x_i - \hat{x}_i)^2\}$$

where $\hat{y}_i$ and $\hat{x}_i$ are the 'best' estimates of y_i and x_i respectively. The estimates of α and β are $\hat{\alpha}$ and $\hat{\beta}$, where

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

$$\hat{\beta} = \frac{(S_y^2 - \phi S_x^2) + \sqrt{(S_y^2 - \phi S_x^2)^2 + 4\phi S_{xy}^2}}{2S_{xy}}$$

An unweighted analysis can be performed. In this case $\phi = 1$ throughout. Such an analysis is built on the assumption that the error variances are equal.

One of the major problems with such an analysis is that the estimates are not invariant to a change of scale, see Cramer (B8). Further,

the distributional properties of the estimators are not obvious.

The use of regression analysis when comparing two, or more, measuring processes is to be strongly discouraged. At best the resulting answers are only indicators of the actual situation being investigated.

We have seen that one of the basic assumptions relating to the model, i.e. the 'independent' variable is measured without error, is not satisfied. Further, the violation of this assumption could cause misleading inferences. A specific numerical example will be given in section 4.5.4.

Basically an inappropriate model is being used to analyse the situation. It has been shown that superficially the regression analysis looks appropriate. To understand more fully the problems with which we are faced, we examine more closely the situations in which comparisons are made between variates measured with error. Having done this, we review some of the standard solutions to the problem of comparison of measuring devices. Some of these solutions are illustrated and compared by means of a specific numerical example.

4.5 Specific Methods of Comparing Numerical Results

4.5.1 Genesis of the Problem

We shall investigate the genesis of the problem. That is, a further description is given of the possible situations which give rise to the comparison of machine results. For the moment we deal with only two machines.

The conditional notation relating to random variables must be introduced before we proceed. The notation Y/X is used to denote the

random variable Y having the 'knowledge of' or 'constraint' X. Thus, if for example we write $Y/X = \alpha + \beta X$ this specifies the relationship of Y for a given X.

To examine how the raw data for comparison is generated, let us follow the description given by Mandansky (1959).

(i). The values X and Y are the true values we wish to observe. We assume that the conditional distribution of Y given X is Gaussian with mean $\alpha + \beta X$ and variance σ^2. For fixed X, irrespective of whether X is a random variable or not, we have

$Y/X = \alpha + \beta X + \phi$

where the term ϕ has a Gaussian distribution with zero mean and variance σ^2. However, we should note that ϕ is not an error of measurement, but represent the inexactness of the association. As we have noted Y and X cannot be observed. We observe, in fact,

$y = Y + \varepsilon$ and $x = X + \delta$

We shall assume that ε and δ are independent Gaussian random variables with zero mean. Further we assume that ε and δ are independent of ϕ and also of X and Y. Our observation on Y for a given X, not x, is

$y = \alpha + \beta X + \phi + \varepsilon$

It is important to realise that the errors ϕ and ε are quite different. Presumably by measuring Y more and more accurately the term ε can be made on average smaller and smaller. However, ϕ is a characteristic of the conditional distribution of Y given X. In some cases, we shall consider ϕ to be zero. If ϕ is non-zero and X is a random variable,

the situation is termed by Mandansky, asymmetric. For although we can investigate

$E(Y/X) = \alpha + \beta X$

the 'inverse' relation $E(X/Y)$ is not obtained by solving the above for X.

(ii). Let Y and X be the true values we wish to observe. We investigate the linear relationship:-

$Y = \alpha + \beta X$

As in (i) above we actually observe

$y = Y + \varepsilon$ and $x = X + \delta$

Again we assume independence as before. Two situations may arise.

(a). X is a random variable - we have a Structural Relationship.

(b). X is fixed but unknown - we have a Functional Relationship.

We note that these are special cases of (i) above, being obtained if ϕ is zero. In Mandansky's terminology we have a symmetric situation. If we investigate

$Y = \alpha + \beta X$

then the relationship

$X = \frac{1}{\beta} Y - \frac{\alpha}{\beta}$ has meaning.

The problems we will encounter will usually be deemed to arise from (ii) either (a) or (b). The question of whether (a) or (b) is applicable is not easily answered. If we take n specimens of blood and measure each by the different methods, we may argue that the bloods are fixed and unknown - suggesting approach (b). However, we could consider that the specimens constitute a random sample from the

population of all bloods, and hence approach (a) would be taken. Usually we will treat this type of situation as arising from (b), (see section 3.3.3).

Before we consider the problems encountered in the estimation of the parameters in (a) or (b), we note an interesting theoretical result in this field. Suppose we wish to investigate the quantities Y and X, but due to error of measurement we observe x_i and y_i where

$$y_i = Y_i + \varepsilon_i \text{ and } x_i = X_i + \delta_i \quad i = 1, 2, 3, \ldots, n$$

In the situation where (y_i, X_i) have a bivariate distribution and there exists the regression

$$y_i = \alpha + \beta X_i + \varepsilon_i$$

where ε_i and X_i are independent we could ask, when does there exist a relationship

$$y_i = \alpha' + \beta' x_i + e_i'$$

Lindley (1947) showed that the conditions for this to happen can be expressed in terms of the cumulants of the distributions of X and δ. Effectively they require these distributions to be members of the same family of distributions. Cochran (1970) has investigated and developed this work. His interest centred on the fact that he felt that Lindley's conditions will rarely apply in practice. This result is interesting, but we shall not develop nor discuss Cochran's ideas. It would appear that Cochran's fears do not have too serious an affect of the estimation process we present.

4.5.2 Problems Encountered in Estimation

I. Structural Relationships

Let us observe the quantities

$$y_i = Y_i + e_i \quad \text{and } x_i = X_i + \delta_i \quad \text{for } i = 1, 2, \ldots\ldots n.$$

Further let us assume that we have the relationship:-

$$Y_i = \alpha + \beta X_i$$

The usual assumptions concerning the random variables are taken to hold. The random variables x_i and y_i are assumed to be jointly Gaussian distributed with the same distribution for all values of i. By substitution we see that we have

$$y_i = \alpha + \beta x_i + (\varepsilon_i - \beta\delta_i) \qquad i = 1, 2, \ldots, n.$$

We may well consider that this is in the form

$$y_i = \alpha + \beta x_i + \delta_i$$

where ϕ_i is an error term. This is obviously true, but we must not mistake it for the similar simple regression model. In this present case, x_i and the error ϕ_i are correlated. This can be seen by considering the following:-

$$\text{Cov}\ (\phi_i, x_i) = \text{Cov}\ (\varepsilon_i - \beta\delta_i, X_i + \delta_i) = -\beta\sigma^2_\delta$$

This contravenes the basic assumptions made in the simple regression analysis. With the usual notation, and under the assumptions given above, we employ Maximum Likelihood estimation to arrive at the following estimates of the parameters:

$$\hat{\mu} = \bar{x}$$

$$\hat{\alpha} + \hat{\beta}\,\hat{\mu} = \bar{y}$$

$$\hat{\sigma}^2_x + \hat{\sigma}^2_\delta = S^2_x$$

$$\hat{\beta}\ \hat{\sigma}^2_x + \hat{\sigma}^2_\varepsilon = S^2_y$$

$$\hat{\beta}\ \hat{\sigma}^2_x = S_{xy}$$

The solution of these equations give the Maximum Likelihood estimates. A solution may not exist, as certain solutions may be inadmissable. Any values are admissable for $\hat{\mu}$, $\hat{\alpha}$ and $\hat{\beta}$, but only positive estimates are admissable for the variances

$\hat{\sigma}^2_x$, $\hat{\sigma}^2_\delta$, $\hat{\sigma}^2_\varepsilon$. (See Kendall and Stuart (B 23)).

The immediate problem is that we have five equations in six unknowns - this is called unidentifiability of the model. Unless we have additional information, we cannot obtain a solution. If we know either of σ^2_ε, σ^2_δ, $\sigma^2_\varepsilon / \sigma^2_\delta$ or both σ^2_ε and σ^2_δ, we may obtain admissible solutions.

Basically we require information concerning the error variances of the measurements. (If both σ^2_ε and σ^2_δ are known, we can maximise the Likelihood Function directly). In practice, we will usually have a reasonably good estimate of σ^2_δ, i.e. the precision of the 'old' or standard method. This will allow the maximum likelihood equations to be solved. If the ratio of the error variances is known, say ϕ, it is easily shown that the estimates obtained are equivalent to those obtained from the 'orthogonal' regression analysis which we have discussed in 4.4. In this analysis the weighted distance of the point (x_i, y_i) is minimised for i = 1, 2, .., n.

One of the major underlying assumptions in the structural relationship model is that x_i and y_i, i = 1, 2, ..., n have a joint Gaussian distribution. It will be unusual if this assumption can be justified in the clinical situations which we discuss. This distributional problem is not found in the functional relationship model which we now discuss.

II Functional Relationship

As before, we observe:-

$$x_i = X_i + \delta_i \quad \text{and} \quad y_i = Y_i + \varepsilon_i$$

Further, as before we have:-

$$Y_i = \alpha + \beta X_i$$

The assumptions concerning independence are as before, but we have

$$E(x_i) = E(X_i) = \mu_i$$

The estimation has been radically altered by introducing a further n parameters to estimate. These parameters are μ_i, i = 1, 2, ..., n. Briefly, we cannot obtain acceptable maximum likelihood estimators unless we again have further information. As before, in the structural case, if we had information relating to the error structure, we could obtain admissible estimators. In the functional situation, it is easy to see why this is so. Suppose that the 'ellipses' shown in figure 4.6, represent the joint confidence regions of X and Y. These regions are 'simultaneous joint confidence intervals' for the true values X and Y. In fact the different levels of confidence would produce a family of concentric ellipses for each point.

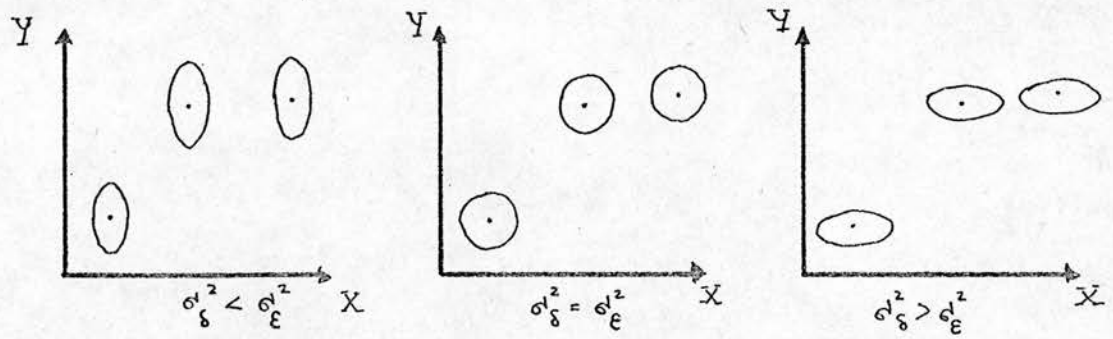


FIGURE 4.6

As we can see, to be able to 'draw' the appropriate line we require to know the eccentricity, which is constant, for these ellipses. The eccentricity in this case is simply the ratio of the error variances of the two observations. However, unlike the previous situation - Structural Relationship, even the knowledge of the ratio of the error variances does not allow for the consistent estimation of all parameters.

As we have indicated a number of solutions have been proposed to extricate us from the estimation problem discussed above. If we have repeat measurements on either or both methods (i.e. either X and/or Y

is replicated), we can estimate the precision(s), and can proceed with the estimation. We shall discuss this later. However, when dealing with blood and especially when comparing more than two methods, the possibility of repeat analysis is reduced - we simply may not have enough blood in each specimen (especially pathological bloods).

Smith (1950) has extended the basic model to describe a more general situation. Again let us consider two machines. In the above, we have assumed that the 'true' values are related by the expression:-

$Y_i = \alpha + \beta X_i \quad i = 1, 2, 3, \ldots, n.$

This relationship is a particular example of the more general:-

$Y_i = f(X_i) \quad i = 1, 2, 3, \ldots, n.$

Smith points out that the 'true' value of the i^{th} specimen is only 'observed' via the machines. Smith suggests that the true value does not exist in reality. Further there is no reason to believe that the 'true' value experienced by a machine is the same as that experienced by some other for the same specimen. The general relationship between 'true' values is too complex to use in practice. Smith therefore suggests that we use the linear relationship which we have discussed.

4.5.3 Practical Solutions to Estimation Problem

We shall now discuss some specific solutions to the analysis problem. These solutions have been proposed by various authors. After presenting the basic outline of each method, we give a numerical example comparing those solutions applicable to comparison of two measuring processes.

(i). The method of Grouping

This approach was suggested by Nair and Shrvastave (1942)

and used by Deming (B12). Let us assume that we are dealing with a Functional Relationship. We will be mainly interested in estimating α and β. This method can be described briefly in the following way.

Consider the observed pairs, (x_i, y_i), $i = 1, 2, .., n$, and order these pairs in ascending order with respect to x_i. Let p_1 and p_3 be proportions such that $p_1 + p_3 < 1$. Group 1, contains the first np_1 of the ordered observations, group 2 contains the next $n(1-p_1-p_2)$ observations and group 3 contains the last np_3 observations. We can estimate β by $\hat{\beta}$ where

$$\hat{\beta} = \frac{\frac{1}{p_1}\sum_{G_1} y_i - \frac{1}{p_3}\sum_{G_3} y_i}{\frac{1}{p_1}\sum_{G_1} x_i - \frac{1}{p_3}\sum_{G_3} x_i}$$

The estimate of α is by $\hat{\alpha}$ where

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

The standard error of $\hat{\beta}$ can be approximated, giving a method of obtaining 'approximated confidence regions'. (See Mandansky). It can be shown that the efficiency of these estimators is a maximum in most situations, when

$$\frac{1}{4} < p_1 = p_3 < \frac{1}{3}$$

(ii). Replicate observations

Let us suppose that we have been able to make replicate measurements on each of n specimens on each of the two machines which we are comparing. Let r_i observations be made on each of the n specimens. Denote by y_{ij}, the j^{th} replicate measurement on the i^{th} specimen by one of the

machines/methods, and x_{ij} be the corresponding measurement on the other machine/method. Thus we can write:-

$y_{ij} = Y_i + \varepsilon_{ij}$ and $x_{ij} = X_i + \delta_{ij}$ $i = 1, 2, .., n$, and $j = 1, 2, \ldots, r_1$.

With the usual assumptions concerning independence and assuming that the random errors ε_{ij} and δ_{ij} have distributions which are Gaussian with mean zero and variances σ_ε^2 and σ_δ^2 respectively. We can present the following Analysis of Variance Table, Table 4.7

Table 4.7	Source	Mean Square	Expected Mean Square
Between	I	$\sum_i r_i(\bar{x}_{i.} - \bar{x}_{..})^2/(n-1)$	$\sigma_\delta^2 + K$
	II	$\sum_i r_i(\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..})/(n-1)$	$\sigma_{\varepsilon\delta} + K\beta$
	III	$\sum_i r_i(\bar{y}_{i.} - \bar{y}_{..})^2/(n-1)$	$\sigma_\varepsilon^2 + K\beta^2$
Within	IV	$\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2/(N-n)$	σ_δ^2
	V	$\sum_i \sum_j (x_{ij} - \bar{x}_{i.})(\bar{y}_{ij} - \bar{y}_{i.})/(N-n)$	$\sigma_{\varepsilon\delta}$
	VI	$\sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2/(N-n)$	σ_ε^2

Where $N = \sum_i r_i$, $K(\text{structural}) = \dfrac{N^2 - \sum r_i^2}{nN-N}$ σ_x^2 and

$K(\text{functional}) = \dfrac{\sum r_i(x_i - \frac{\sum r_i x}{N})^2}{n-1}$

The following estimates of β exist:-

$$\hat{\beta}_1 = \frac{II - V}{I - IV}$$

$$\hat{\beta}_2 = \frac{III - VI}{II - V}$$

$$\hat{\beta}_3 = \sqrt{\frac{III - VI}{I - VI}}$$

$$\hat{\beta}_4 = \frac{(III - \frac{VI}{IV} I) + \{(\frac{VI}{IV} I - III)^2 + 4 \frac{VI}{IV} II^2\}^{\frac{1}{2}}}{2II}$$

We remark that $\hat{\beta}_4$ is the same estimate as we would obtain by carrying out an 'orthogonal' regression analysis. We note with some embarrassment that there are four different estimates. As we shall see later, section 4.5.4, these estimates give different numerical values for the same data.

4.5.4 Numerical example

We now present a specific example, and compare the results obtained by using the analyses which we have presented. The example concerns the comparison of two methods.

Example: We compare two different methods of estimating serum vitamin B_{12}. The two methods are; Lactobacillus leichmanniiand Phadebas. The first mentioned is a microbiological growth method while the second is a radio-isotopic method. Previously we have said that these methods are non-equivalent, but the data we have analysed below has been 'screened' for those pathological specimens which effectively produce non-equivalence. These have been screened using additional medical knowledge and not purely by inspecting the measurements.

Single measurements were made on each specimen by both methods, the data are given in table 4.8

Y:	Phadebas	500	365	240	475	780	470	150	510	690	650
XX:	leichmannii	239	245	110	275	608	442	78	247	562	332

Table 4.8

A cursory inspection reveals that the Y measurements appear to be 'rounded' to give measurements always as a multiple of five; whereas the X measurements appear to be 'rounded' to the nearest unit.

(a). Ordinary Least Squares: Simple Linear Regression.

$\Sigma x = 3138$ $\quad \Sigma xy = 1789180$ $\quad S_x^2 = 278355.6$

$\Sigma y = 4830$ $\quad \Sigma x^2 = 1263060$ $\quad S_y^2 = 344060$

$n = 10$ $\quad \Sigma y^2 = 2676950$ $\quad S_{xy} = 273526$

The product-moment-correlation co-efficient is $r_{xy} = 0.8839$

We calculate the regression of Y on X i.e.

$E(Y/X) = \alpha + \beta X$

From the data above we find that $\hat{\alpha} = 174.660$ and $\hat{\beta} = 0.9826$ and the standard error of $\hat{\beta}$, S_b is 0.1545.

As we have pointed out there seems to be no reason, unlike a true regression case, in determining one regression line rather than the other. So we calculate the estimate of

$E(X/Y) = \alpha' + \beta' Y$

From the data we find that $\hat{\alpha}' = -70.185$ and $\hat{\beta}' = 0.7950$ and the standard error of $\hat{\beta}'$ is 0.1011. As we have mentioned earlier, both of these estimates of 'slope' are underestimates.

(b). 'Orthogonal' Regression Analysis

The 'orthogonal' regression line can be constructed from the data in table 4.8. The estimates of α and β are:-

$$\hat{\beta} = \frac{S_y^2 - S_x^2 + \sqrt{(S_y^2 - S_x^2) + 4 S_{xy}^2}}{2 S_{xy}} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

In this case we have $\hat{\alpha} = 129.35$ and $\hat{\beta} = 1.127$.

If we use the additional information, which was obtained from other data, that the ratio of the precisions is

$\lambda = \sigma_x^2 / \sigma_y^2 = 0.3845$

we can carry out a weighted 'orthogonal' regression analysis. The estimates in this case are:-

$$\hat{\beta} = \frac{\lambda S_y^2 - S_x^2 + \sqrt{(S_y^2 - \lambda S_x^2)^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

In this case we have $\hat{\alpha} = 83.53$ and $\hat{\beta} = 1.273$.

The solutions given in (a) as we have stated before, are not appropriate. The answers obtained from the above have taken into consideration the fact that both values X and Y have error. Thus, on average, we would expect that the weighted estimate should give the 'best' solution, (note that it is the solution which uses the most information), if a linear association held between Y and X. In practice, however, it is likely that $\phi \neq 0$, (see section 4.5.1). In this case what we would like to do is to estimate both precisions and the lack of fit of the linear relationship. To do this an unweighted analysis would be carried out, the total error being 'lack of fit' and the machine imprecisions. This is a complex problem and will not be discussed here.

(c) The method of Grouping

We order the given data with respect to the X measurement, see table 4.9.

Y	150	240	500	365	510	475	650	470	690	780
X	78	110	239	245	247	275	332	442	562	608

We shall consider using different proportions of this data in the groups.

(i). Let $p_1 = p_3 = 3/10$

Then we have $\Sigma_{G1} x = 427$, $\Sigma_{G3} x = 1612$, $\Sigma_{G1} y = 890$, and $\Sigma_{G3} y = 1940$.

From these we calculate the estimates for β and α.

$\hat{\beta} = 0.8861$

$\hat{\alpha} = 204.94$

The standard error of $\hat{\beta}$ is $S_b = 0.31$ (approx.)*

(ii). Let $p_1 = p_3 = 4/10$

Then we have $\Sigma_{G1} x = 672$, $\Sigma_{G3} x = 1944$, $\Sigma_{G1} y = 1255$ and

$\Sigma_{G3} y = 2590.$

From these we calculate the estimates for β and α.

$\hat{\beta} = 1.0495$

$\hat{\alpha} = 153.67$

The standard error of $\hat{\beta}$ is $S_b = 0.53$ (approx.)*

* The standard error of the 'slope' estimate is approximated by constructing a confidence interval, by a method shown by Mandansky, and estimating the standard error from this.

(d). Suppose that two replicate measurements were made on each specimen by both methods. In fact this is how the data previously given was also taken. The previous data was merely the corresponding means of these replicate measurements. The data are given in table 4.10.

TABLE 4.10 on next page

TABLE 4.10

Y		Mean	X		Mean
420	580	500	199	280	239
440	290	365	200	290	245
240	240	240	140	75	110
470	480	475	270	280	275
760	800	780	455	760	608
460	480	470	385	498	442
150	150	150	73	83	78
500	520	510	202	292	247
660	720	690	512	613	562
640	660	650	326	338	332

The Analysis of Variance table, which was given in table 4.7, is given in table 4.11

TABLE 4.11

Source	Mean Square
I	61856.800
II	60783.553
III	76457.776
IV	7100.0
V	1092.0
VI	2730.0

Table 4.12 presents the estimates of 'slope' given previously with the corresponding estimate of the 'intercept'.

	β: 'slope'	α: 'intercept'
$\hat{\beta}_1$	1.0912	140.92
$\hat{\beta}_2$	1.162	118.36
$\hat{\beta}_{33}$	1.125	129.975
$\hat{\beta}_4^*$	1.0534	152.44

* In the discussion following table 4.7 we remarked that the estimate $\hat{\beta}_4$ is equivalent to the orthogonal estimate (using the same data). We note in this example that this is not the case since the 'orthogonal estimate' was determined on single paired readings, whereas $\hat{\beta}_4$ was determined in replicate paired readings.

The example, which we have presented, contained two data sets relating to the comparison of Phadebas, a radio-isotopic method and Lactobacillus leichmannii, a microbiological method of measuring vitamin B_{12}. The first data set comprised single paired readings by both methods. The graphs of the six 'best' fit lines determined are given in figure 4.7. The second data set comprised double (i.e. replicate) paired readings by both methods. The graphs of the four 'best' fit lines determined are given in figure 4.8.

The graphs could have been scaled to give the impression of close agreement. In order to give some idea of the 'agreement' of these lines, we determine the values which would have been forecase for the Phadebas 'normal range' using the lines presented. The 'normal range' used for the Lactobacillus leichmannii method is one presented by Matthews (1962). The forecast values

/contd. over

are presented in table 4.13.

TABLE 4.13

	'Normal Range' 170 (ng/l)	L. leichmannii 970 (ng/l)
Regression, Y on X	342	1128
X on Y	302	1310
Regression, Orthogonal.	321	1223
Weighted Orthogonal.	229	1315
Grouping, 30%	356	1064
40%	332	1172
Replicate, 1.	326	1199
2.	316	1246
3.	321	1221
4.	332	1174

We note immediately, that to carry out such an investigation clinically would require much more data. Further the data sets contain only four low abnormal measurements as judged by the Lactobacillus leichmannii 'normal range'. There are no high abnormal values. Thus, any forecasts which we make using this data will be suspect owing to the concentration of the data in the 'normal - low abnormal range'. However, this illustrates a point which was made previously; that in comparison of machines/methods, the complete clinical range should be covered. The forecasted 'normal range' for Phadebas varies considerably. Thus, although the graphs of the lines may look, at a glance, in agreement, the results of certain inferences can substantially

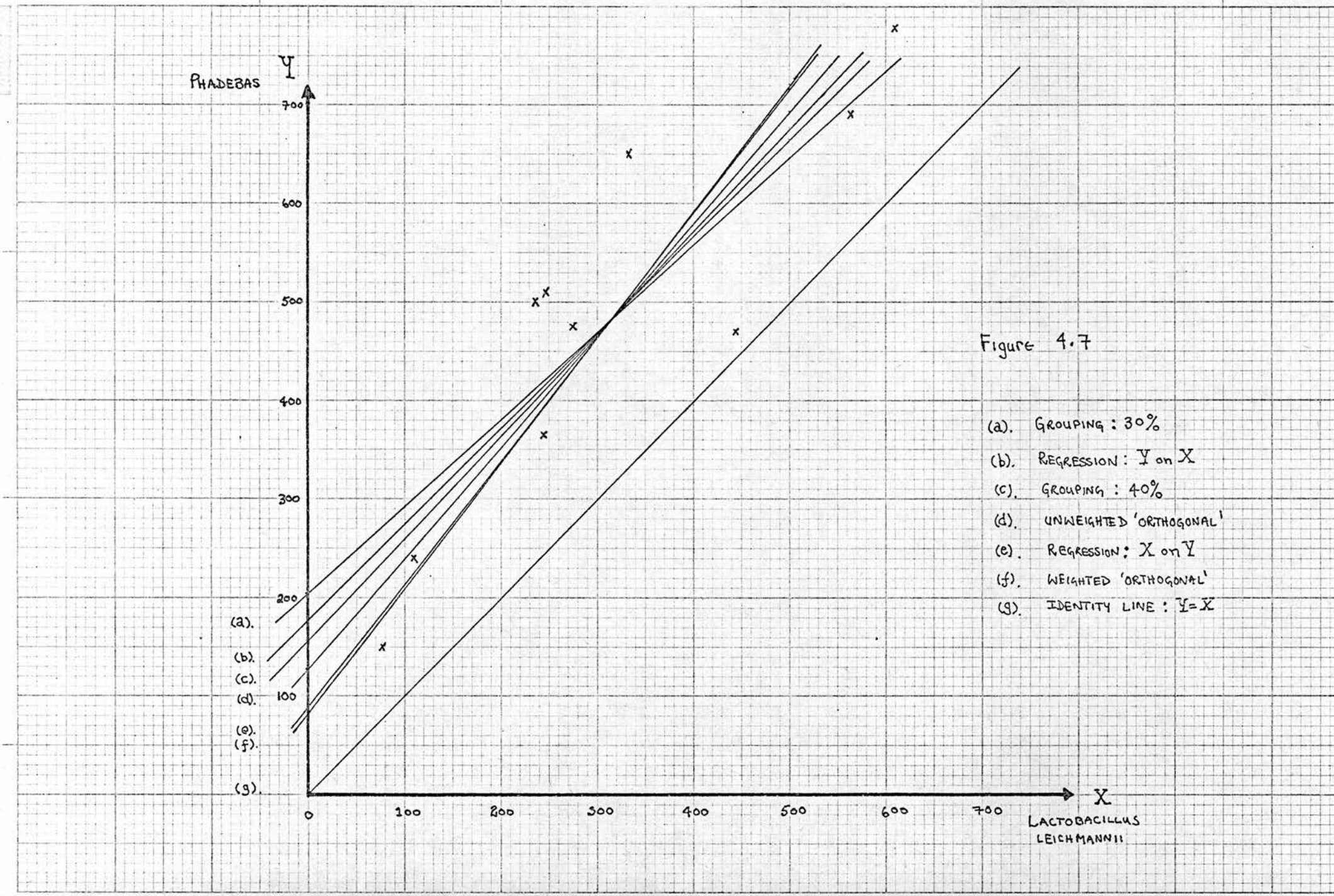


Figure 4.7

(a). GROUPING : 30%

(b). REGRESSION : Y on X

(c). GROUPING : 40%

(d). UNWEIGHTED 'ORTHOGONAL'

(e). REGRESSION : X on Y

(f). WEIGHTED 'ORTHOGONAL'

(g). IDENTITY LINE : $Y = X$

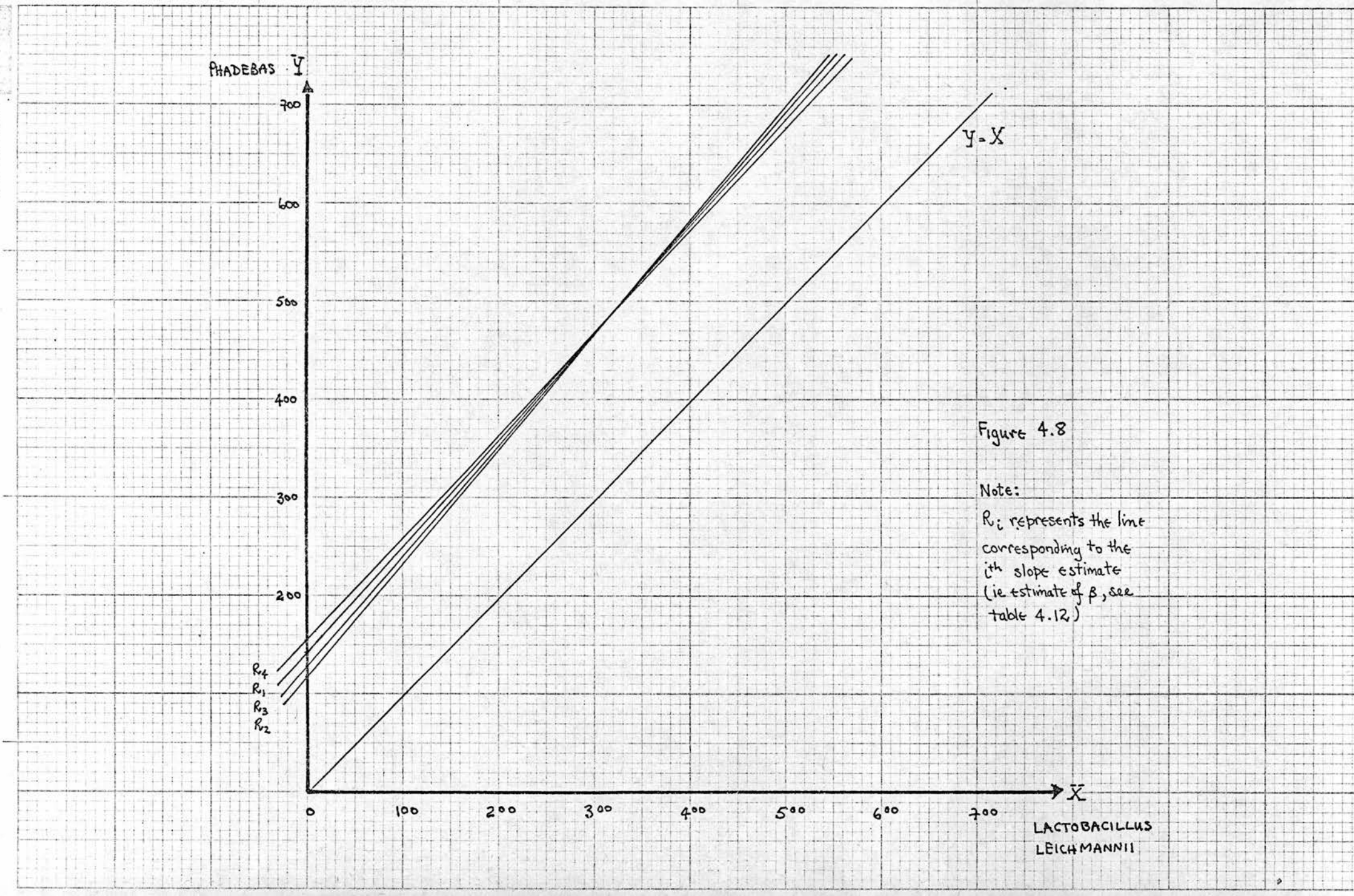


Figure 4.8

Note:

R_i represents the line corresponding to the i^{th} slope estimate (ie estimate of β, see table 4.12.)

differ. For example, the top normal value varies from 1064 ng/l to 1315 ng/l.

Since more information has been used in the replicate estimation, we would expect on average that these would yield the 'best' results. We can categorically dismiss the ordinary regression estimates.

4.6 Two Further Specific Solutions

4.6.1 Introduction

We have discussed some approaches to the problem of comparing machines/methods. Basically these solutions will provide estimates of accuracy and precision of the various methods being compared. Now we shall discuss in detail a further solution to this problem. Grubbs (1948) has shown, given some assumptions, that if three or more methods are compared, biological variability and machine precisions can be separated in the sense that neither the estimate, nor the standard error of the estimate of a machine precision depends on the biological variability. Quite often in the situations we are discussing, it happens that replicate measurements on the same specimen cannot be made, especially if a number of different measuring methods are being compared, the major reason being that there simply is not enough of any particular specimen to replicate more than a few times. Thus, we find that we can either replicate measurements on a single method, or use the specimen once on each method. By replicating a single specimen on a single method we can obtain an estimate of the method's precision. If we replicate a single

measurement obtaining the values $x_1, x_2, x_3, \ldots, x_r$, then an estimate of the method's precision is s^2, where

$$s^2 = \frac{1}{r-1} \sum_{i=1}^{r} (x_i - \bar{x})^2, \text{ where } \bar{x} = \frac{1}{r} \sum_{i=1}^{r} x_i$$

By carrying out similar schemes at different points throughout the clinical range of the parameter we obtain a number of estimates of the method's precision. These estimates allow us to investigate whether the precision of the method is dependent on the value of the clinical parameter being measured. By doing this we can test our assumption of independence between parameter level and machine error.

As we have said Grubbs has shown that by making only one measurement on each method, we can still obtain estimates of the parameters of interest. This will be discussed in the following section. Further, in the following we shall also examine designs in which we amalgamise or hybridise the Grubbs procedure and the replication procedure. We see that if we admit these hybridised designs we may have one of the following situations.

(i). Grubbs - i.e. one measurement per method, each specimen being measured by each method.

(ii). Replication - each specimen is measured on one and only one method, but replicated.

(iii). Grubbs and Replication - one group of specimens are used for a scheme (i) while another is used for a scheme (ii).

(iv). Grubbs with Replications - this allows for each specimen to be replicated on each method.

Grubbs has dealt at length with (i). We shall for completeness, present some of the important results. Case (ii) and (iv) are examples of components of variance situations which are parts of standard

statistical theory. This leaves (iii), which shall be considered in section 4.6.6. We shall make comparisons between these cases giving conditions for one approach to yield more reliable estimates than the others.

4.6.2 Outline of the Basic Models: Grubbs, Replicate

We give outlines of the two basic schemes which will be considered. As we have indicated, these schemes will be called the Grubbs scheme and the Replicate scheme.

(a) Replicate scheme

N machines are to be compared. There are n_k specimens, which can each be replicated r_i times on one machine only. The model is:-

$$y_{ijk} = b_k + x_{ik} + \varepsilon_{ijk} \qquad \begin{matrix} i = 1, 2, \ldots, n_k \text{ (specimens)} \\ j = 1, 2, \ldots, r_i \text{ (replicates)} \\ k = 1, 2, \ldots, N \text{ (machines)} \end{matrix} \qquad 4.6(i)$$

The observation (measurement) y_{ijk}, is the j^{th} replicate which has been made using the k^{th} machine on the i^{th} specimen. The bias of the k machine is b_k. (We note that any of the b_k could be assigned zero - see 4.6.5). The random measurement error is ε_{ijk}.

Two different assumptions concerning the n_k different specimens can be made. Firstly, under the Random Effects Model, we assume the n_k specimens to constitute a random sample from the population. Under the Fixed Effects Model, we assume the specimens are merely a collection not constituting a random sample for the population.

Irrespective of which of these assumptions we choose to make, the

following assumptions are also made. The errors, ε_{ijk}, are independent of each other and of the specimen value x_{ik}. The bias b_k is constant.

(b) Grubbs scheme

N machines are to be compared using n different specimens. Each specimen is measured once only on each of the N machines. The model is:-

$$y_{ik} = b_k + x_i + \varepsilon_{ik} \qquad i = 1, 2, 3, \ldots, n \qquad t = 1, 2, 3, \ldots, N \qquad 4.6(ii)$$

The observation (measurement), y_{ik}, is the value reported for the i^{th} specimen when measured by the k^{th} machine. The bias of the k^{th} machine is b_k and the random error is ε_{ik}.

Again we can have Random or Fixed effects models with the same assumptions concerning the errors as specified in (a).

Comment.

We can have extreme situations under scheme (a). For example, if $n_k = N$, then only one specimen is measured on each of the machines. However, we would usually expect that a 'number' of specimens were replicated on each machine.

Generally, in the situations we are investigating, a large number of specimens can be obtained, but more than likely these specimens will be small in volume. Hence replication is restricted. This also means that the number of machines (methods) to be compared is not large. If we were in a situation in which we had a large number of machines and/or a small volume of blood for each specimen, we could revert to designs which are usually referred to as Balanced Incomplete Block Designs. We shall not discuss these here.

We note that under the Grubbs scheme exactly the same number of measurements are made on each machine. However, under a Replicate scheme, different numbers of specimens could be measured on each machine and further, a different number of replicates could be made on each specimen.

The following constraints could be imposed.

(i). The total number of specimens is n_t. If m specimens are measured on each machine, so that with N machines $Nm = n_t$, we say that we have a 'balance' for machines.

(ii). If the number of replicates on each machine, for each specimen is exactly the same, i.e. $r_1 = r$ $i = 1, 2, \ldots n_r$, then we say that we have a 'balance' in replication.

If both conditions (i) and (ii) are satisfied, we shall say that the design is completely balanced. In general, we shall discuss completely balanced designs. These designs in general are the most efficient i.e. they give the minimum variance estimators over all other designs.

4.6.3 Estimation

(a) Replicate

The replicate model which has been described, can be considered as basically a components of variance model. For the k^{th} machine, assuming we have a balance of replication, we have (see equation 4.6 (i))

$$y_{ijk} = b_k + x_{ik} + \varepsilon_{ijk} \qquad \begin{matrix} i = 1, 2, \ldots, n_k \\ j = 1, \ldots., r. \end{matrix}$$

We define two sums of squares Q_{1k} and Q_{2k} in the following way

$$Q_{1k} = \sum_{ij} (\bar{y}_{i.k} - \bar{y}_{..k})^2 = r \sum_{i} (x_{ik} - \bar{x}_{.k})^2 + r \sum_{i} (\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..})^2 \quad 4.6(iii)$$

$$Q_{2k} = \sum_{ij} (y_{ijk} - \bar{y}_{i.k})^2 = \sum_{ij} (\varepsilon_{ijk} - \bar{\varepsilon}_{i.k})^2 \quad 4.6(iv)$$

Under a Random Effects Model, the expected values of these sums are

$$E(Q_{1k}) = r(n_k-1)\ \sigma_x^2 + (n_k-1)\ \sigma_k^2$$

$$E(Q_{2k}) = n_k(r-1)\ \sigma_k^2$$

where σ_k^2 is the precision of the k^{th} machine and σ_x^2 is the population variance of the clinical parameter being measured.

Under a fixed effects model, the expected values of the sums are

$$E(Q_{1k}) = r \sum_i (x_{ik}-\bar{x}_{.k})^2 + (n_k-1)\ \sigma_k^2$$

$$E(Q_{2k}) = n_k(r-1)\ \sigma_k^2$$

If we denote by $\hat{\sigma}_k^2$ the estimate of the precision of the k^{th} machine, then under either of the models we have the unbiased estimate of the k^m machine precision.

$$\hat{\sigma}_k^2 = \frac{Q_{2k}}{n_k(r-1)} \qquad 4.6(v)$$

Under the Random Effects model, we can obtain $\hat{\sigma}_{xk}^2$ the unbiased estimate of the population variance σ_x^2 from the k^{th} machine. This estimate is:-

$$\hat{\sigma}_{xk}^2 = \frac{1}{r}\left[\frac{1}{n_k-1}\ Q_{1k} - \frac{1}{n_k(r-1)}\ Q_{2k}\right] \qquad 4.6(vi)$$

We find that σ_x^2 is small, the estimate $\hat{\sigma}_{xk}^2$ is sometimes negative. However in the cases which we consider σ_x^2 will usually be larger (much larger) than σ_k^2. Thus negative estimates of variance should not be a problem. However, only rarely will we be interested in estimating σ_x^2.

The variance of these estimates of precision, under the assumption which we have made concerning the errors, can be obtained.

Crump (1951) and Searle (1971) noted that if

$$\hat{\sigma}^2 = a_1 M_1 + a_2 M_2 + \ldots\ldots + a_n M_n$$

where M_i, $i = 1, \ldots\ldots, n$ are mean squares and a_i are constants, then the variance of $\hat{\sigma}^2$ is given by

$$V(\hat{\sigma}^2) = \frac{2a_1^2(E(M_1))^2}{f_1} + \ldots\ldots\ldots + \frac{2a_n^2(E(M_n))^2}{f_n}$$

where f_i, $i = 1, 2, \ldots\ldots, n$ represent the degrees of freedom associated with the i^{th} mean square M_i. Using this result we have

$$V(\hat{\sigma}_k^2) = \frac{2\ \sigma_k^4}{n_k(r-1)} \qquad 4.6(vii)$$

If a balance of replication is not achieved, the result given above are modified in the following way

$$y_{ijk} = b_k + x_{ik} + \varepsilon_{ijk} \qquad i = 1, 2, \ldots\ldots, m$$

$$j = 1, 2, \ldots\ldots, r_i$$

We denote by Q'_{1k} and Q'_{2k} the following sums of squares:-

$$Q'_{1k} = \frac{1}{m-1} \sum^{m} r_i(\bar{y}_{i.k} - \bar{y}_{..k})^2$$

$$Q'_{2k} = \frac{1}{r.-m} \sum^{m} \sum^{r_i} (y_{ijk} - \bar{y}_{i.k})^2 \text{ where } r. = \sum^{m} r_i$$

Irrespective of whether we have a Fixed or Random Effects model the expected value of Q'_{1k} is:-

$$E(Q'_{1k}) = \sigma_k^2$$

If we assume a Random Effects model, we have

$$E(Q'_{2k}) = \sigma_k^2 + r_o\ \sigma_x^2\ ,\ r_o = \frac{1}{m-1}\{r. - \frac{\sum^{m} r_i^2}{r.}\}$$

Crump and Searle, although making typographical errors have given the corresponding variance of the estimate of the precison of the k^{th} machine.

$$V(\hat{\sigma}^2_k) = V(Q'_{1k}) = \frac{2\ \sigma^4}{r, -m}$$

This result reduces to the corresponding result for the completely balanced case. As in other situations in statistics, the completely balanced case will yield the minimum variance estimators for a fixed number of measurements.

Let us assume that the experiments which we discuss are completely balanced. Under a Random Effects model we have N independent estimators of σ^2_x. We could consider that the 'best' combined estimate of σ^2_x would be a weighted estimate of $\hat{\sigma}^2_w$

$$\hat{\sigma}^2_w = \frac{\sum_{k=1}^{N} w_k \hat{\sigma}^2_{xk}}{\sum_{k=1}^{N} w_k}, \quad \text{where } w_k = \left[V(\hat{\sigma}^2_{xk})\right]^{-1}$$

However, this weighted estimate will not usually be much better than a simple average of the estimates, especially since the weights require to be estimated. Further, the weights are strongly influenced by σ^2_x and very little by the machine precisions.

The average estimate $\hat{\sigma}^2_A$ of σ^2_x is:-

$$\hat{\sigma}^2_A = \frac{1}{N} \sum_{k=1}^{N} \hat{\sigma}^2_{xk} \qquad 4.6(viii)$$

This has variance $V(\hat{\sigma}^2_A)$ where:-

$$V(\hat{\sigma}^2_A) = \frac{1}{N^2} \sum_{k=1}^{N} V(\hat{\sigma}^2_{xk}) \qquad 4.6(ix)$$

This can be written in the following way

$$V(\hat{\sigma}^2_A) = \frac{2\ \sigma^4_x}{N(n_k - 1)} + \frac{4\ \sigma^2_x}{N^2(n_r - 1)r} \sum^{N} \sigma^2_j + \frac{2}{r^2 N^2}\left[\frac{1}{n_{k-1}} + \frac{1}{n_k(r-1)}\right] \sum^{N} \sigma^4_j$$

If we can assume that the precisions of the different machines are equal to σ^2, then we have

$$V(\hat{\sigma}_A^2) = \frac{2\sigma_x^4}{N(n_k-1)} + \frac{4\sigma_x^2\sigma^2}{Nr(n_k-1)} + \frac{2}{r^2N}\left[\frac{1}{n_k-1} + \frac{1}{n_k(r-1)}\right]\sigma^4 \quad , \quad 4.6(x)$$

(b) Grubbs

The estimators and corresponding variances of the parameters of interest, were obtained by Grubbs (1948) under the assumption of a Random Effects Model.

$$y_{ik} = b_k + x_i + \epsilon_{ik} \qquad i = 1, 2, \ldots, n$$
$$k = 1, 2, \ldots, N$$

Comparison of two machines

At this stage we reintroduce a notation used in section 4.3.

The following quantities are defined for two variation variables X and Y.

$$S_x^2 = \frac{1}{n-1}\Sigma(x_i-\bar{x})^2, \quad S_y^2 = \frac{1}{n-1}\Sigma(y_i-\bar{y})^2 \text{ and } S_{xy} = \frac{1}{n-1}\Sigma(x_1-\bar{x})(y_i-\bar{y}) .$$

These are simply the unbiased estimates of the population variances and covariance.

Denote by $\hat{\sigma}_k^2$, $k = 1, 2$ and $\hat{\sigma}_x^2$ the estimators of σ_k^2, $k = 1, 2$ and σ_x^2 respectively. Then Grubbs gives the following estimators.

$$\hat{\sigma}_1^2 = S_{y_1}^2 - S_{y_1}S_{y_2} \qquad 4.6(xi)$$

$$\hat{\sigma}_2^2 = S_{y_2}^2 - S_{y_1y_2} \qquad 4.6(xii)$$

$$\hat{\sigma}_x^2 = \frac{1}{4}(S_{y_1+y_2}^2 - S_{y_1-y_2}^2) = S_{y_1y_2} \qquad 4.6(xiii)$$

Under the assumptions of either Random or Fixed Effects model, we have

$$E(\hat{\sigma}_k^2) = \sigma_k^2 \quad k = 1, 2.$$

Under a Random Effects model, we have

$$E(\hat{\sigma}_x^2) = \sigma_x^2 ,$$

whereas under a Fixed effects model, we have

$$E(\hat{\sigma}_x^2) = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2 .$$

Grubbs has shown that under a Random Effects Model,

$$V(\hat{\sigma}_k^2) = \frac{2}{n-1} \sigma_k^4 + \frac{1}{n-1} (\sigma_x^2 \sigma_1^2 + \sigma_x^2 \sigma_2^2 + \sigma_1^2 \sigma_2^2) \qquad 4.6(xiv)$$

$$V(\hat{\sigma}_x^2) = \frac{2}{n-1} \sigma_x^4 + \frac{1}{n-1} (\sigma_x^2 \sigma_1^2 + \sigma_x^2 \sigma_2^2 + \sigma_1^2 \sigma_2^2) \qquad 4.6(xv)$$

Comparison of more than two machines

The results in this section relating to a Random Effects Model were established by Grubbs (1948). We use the same notation as before but extend it to cover the situation of comparisons between more than two machines.

Under the assumption of a Random Effects Model the estimate of σ_x^2 is

$$\hat{\sigma}_x^2 = \frac{2}{N(N-1)} \sum_{1 \leq r < s}^{N} S_{y_r y_s} , \quad N > 2$$

The variance of this estimate is given by

$$V(\hat{\sigma}_x^2) = \frac{2}{n-1} \sigma_x^4 + \frac{1}{n-1} \{ \frac{4}{N^2} \sigma_x^2 \sum_{r=1}^{N} \sigma_r^2 + \frac{4}{N^2 (N-1)^2} \sum_{1 \leq r < s}^{N} \sigma_r^2 \sigma_s^2 \}$$

As we have indicated previously, it is uncommon to want an estimate of σ_x^2. Further, it would be almost impossible to obtain a random sample of specimens from the population which is one of the fundamentals of the analysis.

Under either a Fixed or Random Effects model the estimate of σ_1^2 is

$$\hat{\sigma}_1^2 = S_{y_1}^2 - \frac{2}{N-1} \{ \sum_{r=2}^{N} S_{y_1 y_r} \} + \frac{2}{(N-1)(N-2)} \{ \sum_{2 \leq j < k}^{N} S_{y_i y_k} \}$$

This has the equivalent representation

$$\hat{\sigma}_1^2 = \frac{1}{N-1}\left\{\sum_{s=2}^{N} S_{y_1 - y_r}^2 - \frac{1}{N-2}\sum_{2\leqslant j<k}^{N} S_{y_j - y_k}^2\right\}$$

The variance of the estimate is

$$V(\hat{\sigma}_1^2) = \frac{2}{n-1}\sigma_1^4 + \frac{1}{n-1}\left\{\frac{4}{(N-1)^2}\sum_{r=1}^{N}\sigma_1^2\sigma_r^2 + \frac{4}{(N-1)^2(N-2)^2}\sum_{2\leqslant j<k}^{N}\sigma_j^2\sigma_k^2\right\}$$

Since the numbering of the machine is arbitrary, it is clear that the estimate of the precision of the p^{th} machine and its variance are:-

$$\hat{\sigma}_p^2 = S_{y_p}^2 - \frac{2}{N-1}\left\{\sum_{\substack{r=1\\ r\neq p}}^{N} S_{y_p y_r}\right\} + \frac{2}{(N-1)(N-2)}\sum_{\substack{1\leqslant j<k\\ j\neq p,k\neq p}}^{N} S_{y_i y_k} \quad p = 1,2,\ldots,N.$$

$$V(\hat{\sigma}_p^2) = \frac{2}{n-1}\sigma_p^4 + \frac{1}{n-1}\left\{\frac{4}{(N-1)^2}\sum_{\substack{r=1\\ r\neq p}}^{N}\sigma_p^2\sigma_r^2 +\right.$$

$$\left. + \frac{4}{(N-1)^2(N-2)^2}\sum_{\substack{1\leqslant j<k\\ j\neq p,k\neq p}}^{N}\sigma_j^2\sigma_k^2\right\} \quad p = 1,2,\ldots,N.$$

The point of interest is that provided more than two machines are compared the biological variability and the machine precisions can be completely separated, in the sense that neither the estimate, nor the variance of this estimate of machine precision depends on σ_x^2.

4.6.4 Comparison of Estimators from Models

We compare the estimators obtained by Grubbs and Replicate schemes. Only completely balanced Replicate designs are investigated. We assume that the same total number of measurements are made under the different schemes. (This does not necessarily equate the labour, time nor cost involved).

Let H denote the total number of measurements made under each scheme. Then the following relationships hold.

Grubbs: H = (number of machines) x (number of specimens)

= Nn

Replicate: H = (number of machines) x (number of replications per specimen) x (number of specimens)

= N r n_k .

This leads to the relationship

$$n = rn_k$$

Comparison of Two machines

Estimation of σ_x^2: The model assumed is Random Effects. The simple average estimate of σ_x^2 from the Replicate design is denoted by $\hat{\sigma}^2_{Rx}$. The Grubbs estimate is denoted $\hat{\sigma}^2_{Gx}$. The estimator from the Grubbs scheme is better than the corresponding Replicate estimator if

$$V(\hat{\sigma}^2_{Gx}) < V(\hat{\sigma}^2_{Rx}) \qquad 4.6(xvi)$$

We shall assume that since σ_1^2 and σ_2^2 have similar values, especially in comparison with σ_x^2, that $\sigma_1^2 = \sigma_2^2 = \sigma'^2$. By using equations 4.6(x) and 4.6(xv) 4.6(xvi) reduces to

$$\{(2-r)n_k - 1\}\sigma'^4_x - \frac{2(r-1)}{r}\sigma_x^2\sigma'^2 < \frac{n_k^2r^2(2-r)-(2-r)(r+1)n_k r+1}{r^2 n_k(r-1)}\sigma'^4 \qquad 4.6(xvii)$$

We note immediately that when there are two replications per specimen i.e. r = 2 this condition is always satisfied ($\sigma_x, \sigma' > 0$ and σ'_x is much larger than σ). Where there are more than two replications per specimen i.e. r>2, it will be seen that for all reasonable values of σ'_x and σ' the condition 4.6(xvii) will hold. Hence Grubbs will usually give a better estimate of σ_x^2 than Replicate. This result

should be expected.

Estimation of σ_1^2 and σ_2^2: Denote by $\hat{\sigma}^2_{Rk}$ and $\hat{\sigma}^2_{Gk}$, k = 1,2 the estimates of the machine precisions under the Replicate and Grubbs scheme respectively. Under either Fixed or Random Effects models by using equation 4.6(vii) and 4.6(xiv), the Replicate estimator will be better if

$$V(\hat{\sigma}^2_{Rk}) < V(\hat{\sigma}^2_{Gk}). \quad k = 1,2 \qquad 4.6(xviii)$$

The condition 4.6(xviii) reduces to the inequality

$$\frac{2(n_k-1)}{n_k(r-1)}\sigma_k^4 < \sigma_x^2(\sigma_1^2 + \sigma_2^2) + \sigma_1^2\sigma_2^2 \qquad 4.6(xix)$$

As we have previously stated $\sigma_x^2 >> \sigma_k^2$ k = 1, 2, thus the inequality 4.6(xix) will almost certainly be satisfied. We noted that in many situations we could consider the machine precisions to be equal i.e. $\sigma_1 = \sigma_2 = \sigma$. The inequality 4.6(x x) in this case reduces to:-

$$\tfrac{1}{2}\left[\frac{2(n_k-1)}{n_k(r-1)} - 1\right]\sigma^2 < \sigma_x^2 \qquad 4.6(xx)$$

For all viable designs we note that

$$\frac{2(n_k-1)}{n_k(r-1)} < 2.$$

Thus the inequality (4.6xx) will almost certainly be satisfied. That is we obtain a better estimate of machine precision from the Replicate scheme.

These results are expected. A better estimate of σ_x^2 is obtained from a Grubbs scheme whereas a better estimate of σ_k^2 k = 1,2 is obtained from a Replicate scheme.

Comparison of more than two machines

Estimation of σ_x^2: We discuss for completeness, the estimation

of σ_x^2 under a Random Effects model. Denote $\hat{\sigma}^2_{Gx}$ and $\hat{\sigma}^2_{Rx}$ the estimators of σ_x^2 obtained from the Grubbs and Replicate schemes respectively. The estimator $\hat{\sigma}^2_{Rx}$ is the simple average estimator). For the Replicate estimator to be better, we require

$$V(\hat{\sigma}^2_{Gx}) > V(\hat{\sigma}^2_{Rx}) \qquad 4.6(xxi)$$

If we make the assumption that the precisions are equal, $\sigma_k^2 = \sigma^2$, $k = 1...N$ the inequality 4.6(xxi) reduces to

$$\{N(n-r) - r(n-1)\}\,\sigma_x^4 - 2(r-1)\,\sigma_x^2\,\sigma^2 + $$

$$+ \frac{n(n-r)(r-1) - (N-1)(n-1)^2}{(N-1)n(r-1)}\,\sigma^4 > 0 \qquad 4.6(xxii)$$

For almost all viable designs the left hand side of 4.6(xxii) is negative. That is the condition will not usually be satisfied.

We note that if σ^2 is small in comparison with σ_x^2, then the term in σ^4 can be neglected since the coefficient will not be large. The inequality 4.6(xxii) reduces to

$$\frac{r(n-1) - N(n-r)}{2(r-1)}\,\sigma_x^2 < \sigma^2 \qquad 4.6(xxiii)$$

In the introduction to this section we indicated that we would discuss only those processes for which $\sigma^2 / \sigma_x^2 < 0.1$. From inequality 4.6(xxiii) this would require

$$r(n-1) - N(n-r) < \frac{r-1}{5}$$

which will almost certainly not be satisfied.

Thus as we would expect, the Grubbs scheme usually gives a better estimate of σ_x^2 than a Replicate scheme.

Estimation of σ_k^2, $k = 1,..., N$

We use the same notation as before. Under either the Random or Fixed Effects model, the Grubbs estimate will be better if

$$V(\hat{\sigma}^2_{Rk}) > V(\hat{\sigma}^2_{Gk}) \quad k = 1, 2, \ldots, N \qquad 4.6(xxiv)$$

Now, once more we make the simplification $\sigma^2_k = \sigma^2$ $k = 1, 2, \ldots, N$. If we allow this, it is reasonable to pool the precision estimates to give an estimator of σ^2. Inequality 4.6(xxiv) then becomes.

$$\frac{n(r-1)}{r(n-1)} < 1 + \frac{3 - 2N}{N^2 - N - 1}$$

For viable designs this inequality will almost certainly not be satisfied.

Thus, we have arrived at the same conclusions as before: the better estimate of σ^2_x is from the Grubbs scheme whereas the better estimate of σ^2_k $k = 1, 2, \ldots, N$ is from the Replicate scheme. It should be noted that these are only based on approximate solutions of the inequalities, but in any particular situation we can obtain an exact solution. Further, the conclusions reached agree with our intuition.

4.6.5 Estimation of Bias

In the models we have discussed, we have identified the bias of the k^{th} measuring device by b_k. In fact this is an over-parameterisation. We could define one of the devices to be without bias, say the first device, then $b_1 = 0$. All other biases, b_k, $k = 2, 3, \ldots, N$, represent the relative bias as compared with this chosen unbiased machine. In some cases, where a reference or standard method is available and can be controlled, this will obviously be the choice for zero bias. In other cases, however, the choice may well be arbitrary. In general we can only estimate the relative differences in biases, e.g. $b_2 - b_3$ and not absolute biases e.g. b_2.

The estimation of bias, at least, constant bias, in most preliminary investigations between machines is of secondary importance.

After determinations of the device precisions have been made, the bias can be assessed. However, in any estimation of precision we must allow for a difference in accuracy i.e. bias. Since this is part of the standard procedure for investigating precision, the information concerning the relative accuracies is not unwelcome. It should be obvious that the Grubbs scheme will generally yield more information about bias than the corresponding Replicate scheme. This is intuitive, since Grubbs scheme 'pairs' measurements while the Replicate scheme only allows for a comparison between means.

We shall discuss the comparison of two machines. A simple extension can be invoked to deal with the comparison of more than two machines.

Replicate.

We shall deal with the completely balanced designs. On each of the machines n specimens are measured, each r times. To test for a bias between devices, we rely on a comparison between means.

The hypotheses are:- $H_o: b_1 = b_2$ and $H_1: b_1 \neq b_2$

The test statistic is based on D where

$$D = \frac{1}{n} \sum_i \bar{y}_{i.1} - \frac{1}{n} \sum \bar{y}_{i.2}$$

Under the assumption of a Fixed Effect model, the expected value of D and the variance of D are:-

$$E(D) = b_1 - b_2 + (\bar{x}_{.1} - \bar{x}_{.2})$$

$$V(D) = \frac{\sigma_1^2 + \sigma_2^2}{nr} + \frac{1}{n^2} \sum_i \sum_k (x_{ik} - \bar{x}_{.k})^2$$

Under the assumption of a Random Effects model, the expected value of D and the variance of D are:-

$$E(D) = b_1 - b_2$$

$$V(D) = \frac{2\sigma_x^2}{n} + \frac{\sigma_1^2 + \sigma_2^2}{nr}$$

Under the null hypothesis i.e. H_o: $b_1 = b_2$, the difference $b_1 - b_2$ is zero. We note that with a fixed effect model we have an expected value for D of $\bar{x}_{.1} - \bar{x}_{.2}$. For a Random Effects model the expected value of D is zero. This result is to be expected. Under the assumption of a Random Effects model a random sample of specimens has been taken and randomly assigned to the two methods. Thus when no bias is present the difference of means, as specified by D will 'on average' be zero. However, under the assumption of a fixed effects model the expected value of D is the difference of the means of specimens on both machines. In practice if the specimens are randomly assigned this quantity will be small. (Other balancing out practices could be used to achieve this). Unless we manage of achieve this the test under the Fixed effect model cannot give information unless 'known' specimens are used.

The test statistic under the assumption of a Random Effects model when H_o is true is

$$t_r = \frac{D}{\sqrt{S_D^2}} \quad \text{where } S_D^2 = \frac{1}{n-1}\Sigma(\bar{y}_{i.1} - \bar{y}_{..1})^2 + \frac{1}{n-1}\Sigma(\bar{y}_{i.2} - \bar{y}_{..2})^2$$

The statistic t_r has a t-distribution with 2n-2 degrees of freedom.
Grubbs

Suppose we have n pairs of specimens. The following differences are defined

$$d_i = y_{i1} - y_{i2} \quad i = 1, 2, \ldots, n.$$

In fact, the test is the common 'paired t-test' with hypothesis, as before

H_o; $b_1 = b_2$ and H_1; $b_1 = b_2$.

Under the assumption of a Random Effects model, we have

$$E(\bar{d}) = E(\frac{1}{n} \Sigma\, d_i) = b_1 - b_2$$

$$V(\bar{d}) = \sigma_1^2 + \sigma_2^2$$

Under the assumption of a Fixed Effects model, we have

$$E(\bar{d}) = b_1 - b_2$$

$$V(\bar{d}) = \sigma_1^2 + \sigma_2^2$$

If H_o is true, under both types of models we have $E(\bar{d}) = 0$. Further the test statistic t_g, where

$$t_g = \frac{\bar{d}}{\sqrt{S_d^2/n}} \quad \text{and} \quad S_d^2 = \frac{1}{n-1} \Sigma(d_i - \bar{d})^2$$

has a t-distribution with n-1 degrees of freedom.

4.6.6 Hybrid Schemes

We have already suggested that if we require information concerning machine precisions, variability of the clinical parameter and possible bias, we could do this simultaneously in one hybrid experiment. Often when comparing machines, although not explicitly detailed, a series of experiments are carried out each of which furnishes estimates of the various parameters of interest. What we propose is that with a prior plan, we might design a hybrid scheme which efficiently uses all possible information. Further we would hope to do this in as few measurements as possible.

If we were only interested in estimating the machine precisions, we would proceed with a Replicate design. On the other hand, if we were predominantly interested in estimating bias or the variability of the clinical parameter, we would implement a Grubbs scheme.

In this section we consider combining the information from both types of designs; basically one experiment with two sub-parts. We do this because we wish to use information efficiently. In practice, we could conceivably have a situation in which different volumes of each specimen are available. Thus we may have three or four blood specimens obtained from say a blood bank, of which we have a large quantity e.g. up to one pint of each. The other specimens used for comparison may be actual patient samples of 'excess of patient samples' or blood samples from healthy laboratory staff. In this case we might consider using the 'blood-bank blood' in a Replicate scheme while the other smaller volume specimens are used in a Grubbs scheme, i.e. we have a Hybrid scheme comprising Replicate and Grubbs. We could perhaps consider pooling small volumes of blood to obtain homogeneous batches of bloods of larger volume. However, apart from only giving reliable information concerning machine precision, we would experience some problems caused by the mixing of blood.

We note in passing that the allocation of blood specimens to the different parts of such a hybrid scheme should be at random. (This is a requirement for the statistical analysis.) In practice, it is difficult to obtain large volumes of certain pathological bloods, and hence the precision of the methods must be estimated using small volume samples. The biological variability as reflected by σ_x^2 will not identify any defined population, but in practice this is of no interest.

In the following section we specifically discuss a hybrid scheme which involves two machines.

4.6.7 Hybrid Scheme for Two Machines

The hybrid scheme has two sub-parts. One part follows the Grubbs scheme and the other follows a Replicate scheme, both of which have been described previously.

For the Grubbs part, we make n_g paired measurements on n_g specimens. Thus in total we make $2n_g$ measurements. For the Replicate part we use n_p different specimens for each machine, replicating each r times. Thus the hybrid scheme requires $2n_g + 2n_p r$ measurements to be made.

As we have shown, from both parts of the experiment we can obtain estimates of the machine precisions, σ_1^2 and σ_2^2, and further of σ_x^2, the clinical parameter variability. We can also combine information relating to possible bias. We construct the 'hybrid' estimates by weighting the appropriate estimates from the two sub-parts of the experiment. We have already given these estimates and their corresponding variances. Firstly we note the following general result.

If S_1 and S_2 are two independent estimates of σ^2, with corresponding variances V_1 and V_2, the weighted estimate S, and it's corresponding variance V(S) are

$$S = \frac{V_2 S_1 + V_1 S_2}{V_1 + V_2} \quad \text{and} \quad V(S) = \frac{V_1 V_2}{V_1 + V_2}$$

We introduce the following notation.

$R_A(X, r, n_p)$ denotes the average Replicate estimate σ_x^2, based on n_p specimens per machine, each being replicated r times.

$R_A(E_k, r, n_p)$ denotes the average Replicate estimate of σ_k^2, based on r replicates of n_p specimens.

$G(X, n_g)$ denotes the Grubbs estimate of σ_x^2, based on n_g pairs of measurements.

$G(E_k, n_g)$ denotes the Grubbs estimate of σ_k^2, based on n_g pairs of measurements.

Let the corresponding variances of these estimates be

$$VR_A(X, r, n_p)\ ,\ VR_A(E_k, r, n_p)\ ,\ VG(X, n_g) \text{ and } VG(E_k, n_g).$$

The Hybrid estimator of σ_x^2 is

$$H(X, r, n_p, n_g) = \frac{VR_A(X,r,n_p)\ G(X,n_g) + VG(X,n_g)\ R_A(X,r,n_p)}{VR_A(X,r,n_p) + VG(X,n_g)}$$

The corresponding variance is given by

$$VH(X, r, n_p, n_g) = \frac{VR_A(X,r,n_p)\ VG(X,n_g)}{VR_A(X,r,n_p) + VG(X,n_g)}$$

We have similar expressions for the estimates of the two machine precisions.

$$H(E_k,r,n_p,n_g) = \frac{VR_A(E_k,r,n_p)\ G(E_k,n_g) + VG(E_k,n_g)\ R_A(E_k,r,n_p)}{VR_A(E_k,r,n_p) + VG(E_k,n_g)} \quad k=1,2$$

We note that the expressions for variances and the estimates themselves are functions of n_p, r, and n_g. However they are also functions of the unknown σ_x^2, σ_1^2 and σ_2^2. In practice we use the estimates of these quantities which have been obtained. (We note that often a reasonable estimate is available for the precision of the standard machine.)

We note before proceeding with an example, that there has been a movement in recent years to investigate the construction and properties of optimal designs. Basically such investigations try to provide some method of measuring which of a number of experimental designs is 'best'. Of course 'best' must be defined i.e. what criterion do we use to measure best.

In our situation we could attempt to choose the 'best' design of Replicate, Grubbs and Hybrid. We have seen, however, that analysing Replicate yield the 'best' estimates of precision, Grubbs the 'best' estimates of biological variability and bias and that the Hybrid lies somewhere between these extremes. We do not propose to discuss here methods for choosing the 'Best' Experiment.

4.6.8 Numerical Example

We now give an example of the estimation process for the Hybrid scheme. In doing this, we show how the Grubbs and Replicate estimates are derived.

The following data concerns two methods of estimating measuring the serum vitamin B_{12} using an E.E.L. (blue-green) and Spectra . As we have indicated, the data are collected using a Grubbs scheme and also a Replicate scheme. The data for the Grubbs model are given in table 4.13 and the data for the replicate model are given in tables 4.14 and 4.15.

TABLES 4.13, 4.14 and 4.15 on next page

Grubbs:

Spectra	E.E.L.
250	310
500	580
300	400
460	560
180	230
360	380
350	475
250	220
400	400
280	250
200	217
210	200
390	400
300	233
250	250
600	620

TABLE 4.13

Replicate:

TABLE 4.14

Spectra			
350	400	388	420
150	180	220	220
50	50	31	40
250	250	288	290
325	350	310	280
175	280	280	220
225	200	210	210
400	400	360	440

TABLE 4.15

E.E.L.			
500	450	400	470
325	265	260	270
325	350	350	300
225	175	220	260
275	350	310	330
310	330	385	380
575	580	640	610
500	485	525	560

The data presented under the Grubbs scheme, arose from 16 specimens which were measured on both Spectra and E.E.L. The data presented under the Replicate scheme arose from 16 specimens (different from those in Grubbs scheme). These were randomly allocated to one of the two machines, so that 8 were measured on each machine. Each specimen was measured four times. (This means, using the notation introduced in the last section that $n_g = 16$, $r = 4$ and $n_p = 8$).

We shall obtain the estimates of the machine precisions and the serum B_{12} variability, for the separate schemes and then combine these to give the Hybrid estimates.

Estimation: Grubbs

We let E_i and S_i, $i = 1,2, \ldots, 16$, be the measurements made on E.E.L. and Spectra respectively

$\Sigma S_i = 5280$ $\quad$ $\Sigma E_i = 5725$

$\Sigma S_i^2 = 1948200$ $\quad$ $\Sigma E_i^2 = 2348203$ $\quad$ $\Sigma S_i E_i = 2120350$ $\quad n=16$

$S_S^2 = 13720$ $\quad$ $S_E^2 = 19981.76$ $\quad$ $S_{ES} = 15406.66$

These lead to the following estimates, $\hat{\sigma}_S^2$, $\hat{\sigma}_E^2$ and $\hat{\sigma}_X^2$ of the machine precisions of Spectra and E.E.L. and of the variability of serum B_{12}.

$$\hat{\sigma}_S^2 = 13720 - 15406.66 = -1686.66$$

$$\hat{\sigma}_E^2 = 19981.76 - 15406.66 = 4575.10$$

$$\hat{\sigma}_X^2 = 15406.66$$

The negative estimate of the precision of Spectra is obviously worrying. However such occurrences are not uncommon in variance components situations. Thompson (1963) shows that in such circumstances we should put our estimate of the 'offending' precision to zero. This is

obviously not practically acceptable, but we shall discuss this later. (We note that the product moment correlation coefficient $r_{SE} = 0.9305$).

Estimation: Replicate.

We denote by E_{ij}, the j^{th} replicate on E.E.L. of the i^{th} specimen and by S_{ij} the j^{th} replicate on Spectra of the i^{th} specimen.

TABLE 4.16

$\bar{S}_{i.}$	$\frac{1}{3}\sum_j (S_{ij} - \bar{S}_{i.})^2$
389.50	867.67
192.50	1158.33
42.75	83.58
269.50	507.67
316.25	856.25
238.75	2606.25
211.25	106.25
400.00	1066.67

TABLE 4.17

$\bar{E}_{i.}$	$\frac{1}{3}\sum_j (E_{ij} - \bar{E}_{i.})^2$
455.00	1766.67
280.00	916.67
331.25	572.92
220.00	1216.67
316.25	1022.92
351.25	1372.92
601.25	906.25
517.50	1075.00

We note from tables 4.16 and 4.17 that the variance of the replications does not seem to depend on the mean value of the replicates. This would suggest that the assumption of homogeneous precisions throughout the range of measurement for both methods is reasonable.

In table 4.18 we give the estimates of the biological variation and machine precision. (Formulae discussed in section 4.6.3).

TABLE 4.18 on next page

Spectra	E.E.L.
Q_1 464871.876	376635.875
Q_2 26550.00	21758.00
$\hat{\sigma}^2_S$ = 906.58 $(= 30.11)^2$	$\hat{\sigma}^2_E$ = 1106.25 $(33.26)^2$
$\hat{\sigma}^2_{xS}$ = 13224.64 $(= 115.00)^2$	$\hat{\sigma}^2_{xE}$ = 16326.00 $(= 127.77)^2$

TABLE 4.18

We see that the estimates of the precisions are similar. And further the estimates of the biological variability is similar. The average estimate of biological variability, $\hat{\sigma}^2_{xA}$ is

$$\hat{\sigma}^2_{xA} = \tfrac{1}{2}(13224.64 + 16326.00)$$
$$= 14775.32 = (121.55)^2$$

Using the notation which we have introduced, we have

$R_A(X,4,8) = 14775.32$ $\quad VR_A(X,4,8) = 32272896.94$

$R_A(E,4,8) = 1106.25$ $\quad VR_A(E,4,8) = 101982.42$

$R_A(S,4,8) = 906.58$ $\quad VR_A(S,4,8) = 68490.61$

$G(X,16) = 15406.66$ $\quad VG(X,16) = 34100994.63$

$G(E,16) = 4575.10$ $\quad VG(E,16) = 5243176.99$

$G(S,16) = -1686.66$ $\quad VG(S,16) = 2831614.58$

The Hybrid estimates and their variances can now be determined

$H(X,4,8,16) = 15082.30$ $\quad VH(X,4,8,16) = 16580885.32$

$H(E,4,8,16) = 1172.43$ $\quad VH(E,4,8,16) = 100036.66$

$H(S,4,8,16) = 845.34$ $\quad V(S,4,8,16) = 66873.09$

4.6.9 Maximum likelihood

We indicate, briefly, for the appropriate models, how estimation might be effected by employing the criterion of Maximum Likelihood. The estimates obtained are compared with the estimates previously obtained.

Two different models have been considered for the Grubbs and Replicate designs, i.e. fixed and random effect models. These different models lead to different maximum likelihood solutions.

Replicate.

Fixed effect model.

The mathematical model for two machines in a completely balanced design is:-

$$y_{ijk} = b_k + x_{ik} + \varepsilon_{ijk} \qquad \begin{matrix} i = 1, 2, 3, \ldots, n \\ j = 1, 2, \ldots . r. \quad k = 1, 2 \end{matrix}$$

That is on each machine there are n specimens replicated r times. We consider this to be an experiment comprising r observations which has been undertaken n times. Let α_i and β_i be the 'true' values of the specimens as measured by machine 1 and machine 2 respectively. The log-likelihood function is given by lnL, where

$$\ln L = K - \sum_{i=1}^{n} \frac{r}{2} \log \sigma_1^2 - \sum_{i=1}^{n} \frac{r}{2} \log \sigma_2^2 -$$

$$- \sum_{i=1}^{n} \frac{1}{2\sigma_1^2} \sum_{j=1}^{n} (y_{ij1} - \alpha_i)^2 - \sum_{i=1}^{n} \frac{1}{2\sigma_2^2} \sum_{j=1}^{n} (y_{ij2} - \beta_i)^2 \text{ where k is a constant}$$

Differentiating the log-likelihood function with respect to parameters α_i, β_i, σ_1^2 and σ_2^2 gives the following estimates of these parameters.

$$\hat{\alpha}_i = \frac{1}{r} \sum_{j=1}^{r} y_{ij1} \quad , \quad \hat{\beta}_i = \frac{1}{r} \sum_{j=1}^{r} y_{ij2}$$

$$\hat{\sigma}_1^2 = \frac{1}{nr} \sum^{n} \sum^{n} (y_{ij1} - \hat{\alpha}_i)^2$$

$$\hat{\sigma}_2^2 = \frac{1}{nr} \sum^{n} \sum^{r} (y_{ij2} - \hat{\beta}_i)^2$$

These results are expected. We note that the divisors of the estimates of the precisions differ from the previous results. This is expected, and is the common difference between maximum likelihood and least squares estimators.

Random effect model.

The log-likelihood function cannot be obtained directly as in the above. For a completely balanced design, we consider the mathematical model describing the replicated means.

$$\bar{y}_{i.k} = b_k + x_{ik} + \bar{\varepsilon}_{i.k} \qquad \begin{matrix} i = 1, 2, 3, \ldots, n \\ k = 1, 2 \end{matrix}$$

The assumption made is that $\bar{y}_{i.k}$ has a Gaussian distribution with mean μ_x and variance $\sigma_x^2 + \frac{\sigma_k^2}{r}$.

Let $\theta_k = \sigma_x^2 + \frac{\sigma_k^2}{r}$ $k = 1, 2$

The log-likelihood function is given by:-

$$\ln L = K - \frac{n}{2} \ln\theta_1 - \frac{n}{2} \ln \theta_2 - \frac{1}{2\theta_1} \sum_{i=1}^{n} (\bar{y}_{i.1} - \mu_x)^2 - \frac{1}{2\theta_2} \sum_{i=1}^{n} (\bar{y}_{i-2} - \mu_x)^2$$

Differentiating with respect to μ_x the estimate $\hat{\mu}_x$ of μ_x is obtained. However, maximum likelihood estimates for σ_x^2, σ_1^2 and σ_2^2 cannot be obtained simultaneously. Maximum likelihood estimates are available for θ_1 and θ_2.

$$\hat{\mu}_x = \frac{\hat{\theta}_2 \bar{y}_{..1} + \hat{\theta}_1 \bar{y}_{..2}}{\hat{\theta}_1 + \hat{\theta}_2}$$

$$\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} (\bar{y}_{i.k} - \hat{\mu}_k)^2 \quad k = 1, 2.$$

We could use the estimates of σ_1^2 and σ_2^2 derived in the above to separate the joint estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$. These are intuitive and reasonable estimates of the precisions. If this is done, two estimates of σ_x^2 are obtained.

$$\hat{\sigma}^2_{x1} = \frac{1}{n} \sum_{i=1}^{n} (\bar{y}_{i.1} - \hat{\mu}_x)^2 - \frac{1}{nr^2} \sum^{n} \sum^{r} (y_{ij1} - \hat{\alpha}_1)^2$$

$$\hat{\sigma}^2_{x2} = \frac{1}{n} \sum^{n} (\bar{y}_{i.1} - \hat{\mu}_x)^2 - \frac{1}{nr^2} \sum^{n} \sum^{r} (y_{ij2} - \hat{\beta}_i)^2$$

These estimates can be compared with the estimators which have been derived in section 4.6.3. In this case, it would be reasonable to construct a pooled estimate for σ_x^2.

Grubbs

Fixed effect model

There is no obvious way to obtain estimates of the machine precisions under the assumption of a fixed effect model. The problem is that the precisions are confounded and even by considering the sum and difference of observations only give a joint estimate of $\sigma_1^2 + \sigma_2^2$.

Random effect model.

Denote by $\underset{\sim}{y}_i'$ the vector $(y_{i1}, y_{i2})'$. The mean vector of $\underset{\sim}{y}_i'$ is denoted by $\underset{\sim}{\mu}_x'$ and the covariance matrix (dispersion matrix) by Σ.

An equivalent way of expressing our previous assumptions for the random effect model is to assume that $\underset{\sim}{y}_i$ has a bivariate Gaussian distribution with mean vector $\underset{\sim}{\mu}_x' = (\mu_x, \mu_x)'$ and dispersion matrix:-

$$\Sigma = \begin{bmatrix} \sigma_x^2 + \sigma_1^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \sigma_2^2 \end{bmatrix}$$

A random sample of n pairs is obtained and the corresponding log-likelihood function is:-

$$\ln L = k - \frac{n}{2} \ln \left[(\sigma_1^2 + \sigma_2^2)\ \sigma_x^2 + \sigma_1^2\ \sigma_2^2\right] - \frac{Y_1(\sigma_x^2 + \sigma_2^2) - 2\ \sigma_x^2 Y_{12} + Y_2(\sigma_x^2 + \sigma_1^2)}{2\left[(\sigma_1^2 + \sigma_2^2)\ \sigma_x^2 + \sigma_1^2\ \sigma_2^2\right]}$$

where k is constant, $Y_j = \sum_{i=1}^{n} (y_{ij} - \mu_x)^2$ $j = 1,2$ and

$$Y_{12} = \sum^{n} (y_{i1} - \mu_x)(y_{i2} - \mu_x)$$

Differentiating the log-likelihood function with respect to μ_x results in the estimate $\hat{\mu}_x$

$$\hat{\mu}_x = \frac{\hat{\sigma}_1^2\ \overline{Y}_2 + \hat{\sigma}_2^2\ \overline{Y}_1}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} \quad \text{where } \overline{Y}_k = \frac{1}{n} \sum^{n} y_{ik}, k = 1,2\ .$$

This is a weighted average, which is expected. The weighting is in relation to the precision of the measuring devices.

Differentiating the log-likelihood function with respect to σ_{x1}^2, σ_1^2 and σ_2^2 gives the following estimators.

$$\hat{\sigma}_x^2 = Y_{12}/_n, \quad \hat{\sigma}_k^2 = \frac{1}{n}(Y_k - Y_{12}), k = 1, 2$$

These estimators suffer from the usual 'bias' of Maximum likelihood estimators. With the usual change of divisors the previous estimators are obtained.

4.6.10 Tests and Related Problems

In general we require to find estimates for the precision and accuracy of the machines which we are comparing. Grubbs (1948) has defined these terms as follows.

Precision is defined as a measure of the variation in the errors of measurement. Thus, precision is sometimes referred to as the reproducibility of the machine in question.

Accuracy is the closeness of the measurement to the true value of the material being measured.

These definitions have been subsequently adopted by other authors. We see that these definitions are the same as those given in the previous chapter. In almost all circumstances for an efficient measuring procedure we require the precisions of the measuring devices to be at least an order of magnitude less than the variability of the material being measured.

We have noted that in the construction of estimates in the Grubbs case we found similarities to variance components estimates. Thus, it is possible for estimates of precision to be negative. Thompson (1963) considers Restricted Maximum Likelihood estimation which coincides with Grubbs estimates when these are positive. When any of these estimates are negative we replace them by zero. This is intuitively appealing. We can assume that one of two situations have occurred when we obtain negative estimates. Either the model we have used is inadequate or the model is adequate, but small random perturbations have occurred resulting in the negative estimates. (Hanumara (1975) presents non-negative estimates of variance in the situation in which three measuring instruments are compared.)

In the previous sections we have constructed estimators for the various parameters of interest. We have shown that if we wish to have information concerning all the parameters of interest we should, on average, be better off if we use a hybrid scheme. If we were interested in specific parameters we should opt for one of Grubbs or Replicate. Hypothesis Testing under a Replicate scheme is standard, being given in most reviews of analysis of variance. However, testing under a Grubbs scheme is not a part of standard theory. Many of the tests which have been given have appeared within the last seven years. We shall give a brief review of this work in this section.

The situation described by the Grubbs scheme can be modelled by one of the following two equivalent models.

(i). $y_{ij} = x_i + \beta_j + e_{ij}$ $i = 1, 2, \ldots, n.$ and $j = 1, 2, \ldots, N.$

We assume that x_i is a Gaussian random variable with mean and variance, μ_x and σ_x^2.

We assume that e_{ij} is a Gaussian random variable with mean and variance, 0 and σ_j^2. β_j is the bias associated with the j^{th} machine.

(ii). $y_{ij} = x_i + e_{ij}$ $i = 1, 2, \ldots, n.$ and $j = 1, 2, \ldots, N$

We assume that x_i is a Gaussian random variable with mean and variance, μ and σ_x^2

We assume that e_{ij} is a Gaussian random variable with mean and variance β_j and σ_j^2

The representation given in (i) has been used by Hahn and Nelson (1970), Grubbs (1973), Jaech (1971) and Maxwell (1974). The representation (ii)

has been used by Grubbs (1948) and Thompson (1963). Comparisons have been made by Gaylor (1956) and Russell and Bradley (1958) who gave the alternative model

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where μ is an overall mean, $E(\alpha_1^2)$ = the variability of the clinical parameter and β_j is the bias of the j^{th} machine. The assumptions concerning ε_{ij} are the same as in (i). This we see is just the usual two-way-classification model.

Commonly accepted assumptions

The commonly accepted assumptions concerning the model given in (i) are as we have indicated: Gaussian and independence. We let

Variance of $x_i = \sigma_x^2$ = the variability of the clinical parameter

and Variance of $\varepsilon_{ij} = \sigma_j^2$ = Precision of the j^{th} machine.

Grubbs (1948) accepts that no assumptions are necessary to estimate σ_x^2 and the σ_j^2 using model (i). However with the Gaussian assumption for x_i and e_{ij}, and independence between errors and specimen true values, he shows that the estimators derived are functions of sufficient statistics. Further in the case where N = 2 these estimates are Maximum Likelihood estimates. With Gaussian assumptions for the model Gaylor showed that the estimates are equivalent to variance component estimates.

Assumptions of independence are required to be made before we can use much of the standard statistical theory. Grubbs (1948) points out that this assumption may not be justified. Commonly errors are related to the specimen's value and often increase as this value increases. Grubbs restricts the variability of the material he measures, i.e. restricts the specimens to be sampled from the clinical

normal range. By doing so we hope that the assumption of independence is better justified, and of course we also make the error variances more homogeneous.

Grubbs (1948) noted at the end of the paper that although he had not presented any tests for the comparison of machine precisions, nor for that matter bias, he thought that 'certain rough or approximate tests may be evident to the reader'. A warning is given concerning the dependence between the estimates and the difficulties in giving the degrees of freedom, e.g. for the estimate of σ_j^2 when three machines are being compared, the degrees of freedom are given as approximately $2(n-1)/5$.

In a general paper, concerning one-way variances in a two-way classification, Russell and Bradley (1958) investigated tests and estimates for the equivalent model given previously. They approach the problem as an extension of an analysis of variance problem and looked for a 'least-squares solution'. They note that when we are comparing two machines, a simple but inconsistent solution to the 'normal equations' exists. They also show that the Maximum Likelihood estimate of σ_j^2 is

$$\hat{\sigma}_j^2 = \frac{1}{n} \sum_i \left[y_{ij} - y_{.j} - \frac{\sum_{q=1}^{N} (y_{iq} - y_{.q}) / \hat{\sigma}_q^2}{\sum \frac{1}{\hat{\sigma}_q^2}} \right]^2 \qquad j = 1,2,3,\ldots,N$$

When we have two machines for comparison, iterative procedures are needed to obtain such estimates. Such procedures have been shown to converge. Jaech (1971) has also discussed this problem. Russell and Bradley indicate that iterative procedures will converge for $N > 2$ only if we put $\hat{\sigma}_j^2 = 0$ for some j. Russell and Bradley conclude that

the reason for the above 'not working' is that we have used direct methods on the Likelihood Function.

Now if $\sigma_j^2 = \sigma^2$ for all j, i.e. all machines had the same precision, then σ^2 would be estimated in the usual way as for a two-way classification. Even if $\sigma_j^2 \neq \sigma^2$, this estimate depends on (n-1)(r-1) 'linear contrasts' formed from the original observations, and do not depend on μ_i. or β_j.. This is not a Maximum Likelihood estimator with respect to the usual likelihood function, but is a maximum likelihood estimator with respect to the likelihood function of any set of (n-1)(r-1) linear independent error contrasts. Thus by transforming the data and restricting our attention to a set pf (n-1)(r-1) error contrasts, the problem should be circumvented. By doing this, Russell and Bradley achieve the same estimators for the error variances as did Grubbs. Further they were able to deduce the allied distributional and test theory. However, they concluded that in the general case the associated distributions were complex. They did produce large sample tests for homogeneity of variances with N=3. Further, an exact test for homogeneity of variances knowing that all but one specified variance may be different was also given for small samples and unrestricted N and n.

Thompson (1963) investigates the relative precision of a measuring process. This he defines to be, in our notation, $\Delta_i = \frac{\sigma_x}{\sigma_i}$ the ratio of the clinical parameter variability to the machine precision. As mentioned previously, a 'rule of thumb' requires $\Delta_i > 10$ for an efficient measuring procedure. Under the assumption of normality, inferential statements can be made concerning Δ_i. Thompson also gives a method of constructing confidence intervals for the actual variance estimates.

The first paper to discuss in a more specified sense, the testing of precisions was by Hahn and Nelson (1970). This paper discussed the two machine situation where, for one of the machines a replicate measurement was made. (An extension is considered where although one measurement is made on one of the devices, replicate measurements are made on the other.)

The model presented is

$$x_i = A_i + M_x + e_{xi}$$

$$y_i = A_i + M_y + e_{yi} \qquad i = 1, 2, \ldots, n$$

$$y'_1 = A_i + M_y + e'_{yi}$$

The quantities on which the analysis is based are

$$u_i = x_i - \tfrac{1}{2}(y_i + y'_i) \text{ and } v_i = y_i - y'_i \qquad i = 1, 2, 3, \ldots, n$$

The usual assumption of Gaussian variates and independence are made. Using these a t-test is constructed to test for accuracy differences and an F-test is constructed to test for equality of machine precisions. An extension is possible to cover situations in which the measurements made on different machines could have been made at different times. In this case time effects and accuracy are confounded - except that we can test for differences in time effects for the replicate readings.

Maloney and Rastogi (1970) confined themselves to tests for two instruments, using single measurements. The tests based on the test for comparing marginal variances in a bivariate distribution with non-zero correlation, which was introduced by Pitman (1939) and Morgan (1939). The hypotheses, concerning the precisions of the two machines, take the form:

$$H_o: \sigma_1^2 = \sigma_2^2 \text{ against the alternative } H_1: \sigma_1^2 \neq \sigma_2^2$$

If we are interested in testing approximately that one of the machines is precise i.e. $\sigma_i^2 = 0$ then by using the likelihood ratio

$$\lambda = \frac{S_x^2 S_y^2 - S_{xy}^2}{\left[S_x^2(S_x^2 + S_y^2 - 2 S_{xy})\right]^{n/2}}$$

In cases in which n is large, we can use the fact that $-2\ln\lambda$ has an approximate χ^2 distribution with one degree of freedom.

Further tests for Grubbs estimates were given by Jaech (1971). He considered the case in which two machines had stated precisions and, it is required that we judge whether these statements are valid or not. This situation will arise in practice where a manufacturer states the precision (expected) from the device. Once again Jaech bases the tests on the likelihood ratio for large n. If the manufacturers state that the precision of their machines should be σ_{10}^2 and σ_{20}^2, these tests allow for hypotheses of the type

H_o: $\sigma_1^2 = \sigma_{10}^2$

H_o: $\sigma_1^2 = \sigma_{10}^2$ and $\sigma_2^2 = \sigma_{20}^2$

H_o: $\sigma_1^2 = k\ \sigma_2^2$, k specified.

Shukla (1973) has slightly extented this work. He based the approach on likelihood ratio tests.

Grubbs (1973) has suggested that the Pitman-Morgan test in the case of two machines is insensitive to detecting differences in precisions excepting for large n. This is due to the influence of the clinical parameter variability i.e. σ_x^2. Grubbs suggests using three machines - two standard machines and one test machine. (We note that in section 4.6.3 it was shown that three or more machines were required

to produce estimates of precision not dependent on σ_x^2). The quantities on which the analysis is based are the differences of the three measurements in pairs. Essentially we are considering that two machines are standard, say 1 and 2. Thus we could be interested in whether $\sigma_1^2 = \sigma_2^2$ and testing the average precision of the standard machines with the third, i.e. the test machine. Also interest is directed towards any bias between the standard machines, i.e. testing for $\beta_1 - \beta_2$. This could lead us to considering the possibility of a difference between the average standard machine reading and the third reading, that is the reading from the test machine. The test suggested by Grubbs are based on the t-test, F-test and a modified Pitman-Morgan test. Grubbs notes that in this situation the significance tests depend only on the errors of measurement being Gaussian distributed.

Maxwell (1974) investigates more fully the situation in which we compare two machines, making a single measurement on one machine but replicate measurements on the other. In this particular case, the two machines are thought of as being in two different laboratories. Confidence intervals are given for the precision of the second machine and the ratio of the precision of the two machines.

Jaech (1976) has investigated tests concerning Grubbs estimators in situations of more than two machines. The tests are large sample tests, based on a likelihood ratio type statistic, which is constructed using least squares estimates. Simultation studies are included.

Example.

We use the data given in the previous example concerning the comparison of E.E.L. (blue-green) and Spectra methods of measuring serum vitamin B_{12}. Let σ_E^2 and σ_S^2 denote the precision of E.E.L.

and Spectra respectively.

Consider a test of hypothesis concerning the equality of the precisions.

$$H_o: \sigma_S^2 = \sigma_E^2$$

$$H_1 \quad \sigma_S^2 \neq \sigma_E^2$$

To test this, we use the test procedure suggested by Maloney and Rastogi (1970), which is based on the Pitman-Morgan test. The same notation is used as in section 4.6.8 i.e. denote by E_i and S_i the measurements made on the i^{th} patient specimen by E.E.L. and Spectra respectively. The sums and differences p_i and d_i we computed:-

$$p_i = E_i + S_i \text{ and } d_i = E_i - S_i \qquad i = 1, 2, \ldots, n.$$

The product moment correlation coefficient r_{dp} is computed for these data. Under the Null hypotheses, H_o, the test statistic t, where

$$t = \frac{r_{dp}\sqrt{n-2}}{\sqrt{1 - r_{dp}^2}}$$

has a t-distribution with n-2 degrees of freedom.

From the data presented in table 4.13, we find that $r_{dp} = 0.44705$ and so t = 1.87. For the student t-distribution with 14 degrees of freedom, the 5% point is 2.145 and the 10% point is 1.761.

Although in this case there is an indication that a difference between precisions exists, in practice much more data would be available; and would be required.

4.6.11 Conclusions

We have given a detailed discussion of the estimation of machine precisions and other quantities of interest under a Grubbs scheme.

Although many authors have investigated schemes which may be classified under the general heading of Grubbs schemes, some criticism can be levelled against such analysis. Basically these criticisms centre on the fact that no allowance is made in the models for a proportional error difference between machines. However we may suggest the following counter argument. We would expect that any machine which has been marketed, has undergone extensive tests before coming on to the general market. If this is so, we might expect that proportionality differences would not be expected, or at least only infrequently.

4.7 Instrumental Variables

It will be obvious that the Replicated design can be extended to situations in which three or more machines are compared. We have already described how the Grubbs analysis is extended in such a situation. To close this discussion on the comparison of machines we describe briefly an alternative approach proposed by Barnett (1969). The investigation concerns the comparison of three or more machines simultaneously. The basis of the analysis depends on 'instrumental variables'.

Instrumental variables are simply a set of variables 'closely' related to the random variables of interest. Barnett uses an instrumental variable which is linearly related to the random variables of interest. In practice, this may be difficult to achieve. However, Barnett supposes that a reference method exists which is linearly related to all other measurements to be made. We shall give a brief description here.

Suppose that we have p random variables X_o, X_1, X_2,...,X_{p-1}, which satisfy structural relationships

$$X_i = \alpha_i + \beta_i X_0 \quad i = 1, 2, \ldots, p-1.$$

It is not possible to observe the X_i, but instead we observe x_i where

$$x_i = X_i + \varepsilon_i \quad i = 1, 2, \ldots, p-1.$$

The ε_i are random errors, which are Gaussian distributed with zero mean and variance σ_i^2. $i = 0, 1, 2, \ldots, p-1$. Let us assume that X_0 is Gaussian distributed with mean μ and variance σ^2. Further we assume that the X_i and ε_i are uncorrelated.

In practice, Barnett says that we could consider that the random variable x_o is a measurement from a reference machine/method. (The assumption of linearity, in our case, between a number of methods, may not be well founded). By convention we put $\alpha_o = 0$ and $\beta_o = 1$.

This description is more general than the previous ones. In the previous cases we considered only two machines/methods, now we have admitted essentially any number. In our previous case, $p = 2$, i.e. two machines/methods we found that the problem of undentifiability required us to obtain additional information, before we could carry out the estimation of the parameters. However, when $p > 3$ we no longer experience this problem.

Barnett deals with the case when $p = 3$, developing estimators and the corresponding standard errors. He develops this work in situations when $p > 3$.

An alternative way of considering this approach is by comparing it with the corresponding standard Factor Analysis model.

Let us suppose that we have p machines/methods to compare and further suppose that the true values as presented to the machine are

z_i, i = 1,2,...,p. Let k of these be linearly independent and suppose we express the remaining p-k as linear functions of the others. If $E(z_i) = \mu_i$, then

$$z_i = \mu_i + \sum_{j=p-k+1}^{p} \alpha_{ij}(z_j - \mu_i), \quad i = 1, 2, \ldots, p-k$$

Let us assume that the independent random variables, $z_{p-k+1}, \ldots, z_p$, follow a multivariate Gaussian distribution, non-singular, with mean vector, μ, and dispersion matrix Σ. Now the measurements reported by the machine are $x_1, x_2, \ldots, x_p$ and these are

$$x_i = z_i + e_i, \quad i = 1, 2, 3, \ldots, p.$$

That is the observations have errors, and these errors are independent of the $z_1, z_2, \ldots, z_p$. Further we assume that they are Gaussian distributed with mean zero and variance χ_i. We can now write this as:-

$$x_i = \mu_i + \sum_{r=1}^{k} \lambda_{ir} f_r + e_i \qquad i = 1, 2, 3, \ldots, p.$$

where the f_i, i = 1, 2, ..., k have standard Gaussian distribution. This is the basic factor analysis model. We note, however, that this approach can only be used if we are comparing more than two machines. As we have indicated, the use of instrumental variables in haematological comparisons will not usually be possible in practice. The major reason is that the assumption concerning the linearity between a single method and all other 'test' methods will not be justified.

4.8 Non-Equivalent machines/methods

In the introduction to this chapter, we identified a problem concerning the comparison of measuring methods. Two different types of comparison could be undertaken. Firstly, there was comparison between equivalent machines/methods. Indices and methods for such comparisons have been dealt with in the sections 4.2-4.6. Secondly,

there was the comparison of non-equivalent machines/methods. We discuss only briefly in this section the problems relating to the comparison of non-equivalent machines/methods.

It was noted that Smith (1950) mentioned the problem of 'non-equivalent' measuring processes. He outlined the basic problem as follows.

'Consider that the problem is to compare two instruments for measuring a character which can only be observed via these instruments, so that there are no theoretically absolute values against which the instruments can be calibrated; the definition of the character being in fact determined by the preferred method of measurement. In these circumstances it would be necessary to consider both accuracy and scale jointly.'

It will be seen that this type of situation occurs commonly in clinical situations, and specifically in haematology. In particular, in biological assay, since no assay is specific in nature different assay procedures could be considered to be non-equivalent measuring processes. Let us consider the situation in the following way.

We consider the comparison of two machines/methods. Let the first machine produce a measurement y_{i1} for the i^{th} specimen, where

$$y_{i1} = x_{i1} + \varepsilon_{i1} \quad i = 1, 2, 3, \ldots, n.$$

The 'true' value of the i^{th} specimen, as measured by this method is x_{i1} and the measurement error is ε_{i1}. Let the corresponding measurement reported by the second machine be,

$$y_{i2} = x_{i2} + \varepsilon_{i2} \quad i = 1, 2, 3, \ldots, n.$$

In this case x_{i2} is the 'true' value of the i^{th} specimen by this method of determination. In the case of exactly equivalent machines,

we have $x_{i1} = x_{i2}$, $i = 1, 2, \ldots, n$. We may also include under the heading of equivalent processes, as we have done in the preceding sections, the case in which $x_{i1} = \alpha + \beta x_{i2}$ $i = 1, 2, \ldots, n$. In fact the most general equivalent situation would admit a relationship of the form $x_{i1} = f(x_{i2})$, where $f(.)$ is some function linear or non-linear.

If the processes are not equivalent, we represent the situation by,

$$y_{i1} = x_{i1} + \varepsilon_{i1}$$

$$y_{i2} = f(x_{i1}, \theta_{ij}, j = 1, 2, \ldots, s) + \varepsilon_{i2} \quad i = 1, 2, 3, \ldots, n.$$

The 'true' value for the first machine is x_{i1}, whereas the 'true' value for the second machine is $f(x_{i1}, \theta_{ij}, j = 1, 2, \ldots, s.)$. The 'true' value for the second machine is not only a function of the corresponding 'true' value of the first machine but also of the factors θ_{ij}, $j = 1, 2, \ldots, s$. These are factors which affect the second machine's results. In the model presented any bias is measured with respect to the first process, and any affecting factors are measured with respect to the first process. Let us consider two examples to illustrate these points.

We have mentioned in the introductory chapter, the evaluation undertaken by Widmark and Weschler (1970) on the Technicon Auto-Analyser when counting platelets. The evaluation indicated that speciments (patients) experiencing one of the following factors (see table 4.19) produced 'false' readings.

TABLE 4.19 on next page

TABLE 4.19

Platelet count too high.	High leucocyte count, Nucleated erythrocytes, Howell-Jolly bodies, Malaria parasites, Abnormal high serum proteins, Erythrocyte auto-antibodies.
Platelet count too low.	Platelet auto-antibodies, Chemotherapy

We should note that the concept of 'too high', 'too low' and 'false' depend on another independent, and perhaps absolute, scale of measurement. In comparing these platelet estimations against some other method of platelet counting, the factors listed in the above table would be represented by one each of the θ_{ij}.

Usually such a detailed investigation of what must be regarded as pathological clinical specimens is not carried out at the evaluation stage, although patholgical specimens are investigated. The anomolies are usually found after some preliminary period of use. The effects which we have indicated, will commonly only be identified by laboratories who investigate inconsistent results, and have confidence in their reliable clinical findings.

Let us consider a second example. We compare two methods for measuring serum vitamin B_{12}. The first of these methods uses radioisotopic tagging, while the second employs microbiological growth. Several procedures have been suggested for both methods.

Dacie and Lewis, 'Practical Haematology' 5th Ed. have given some examples of clinical states and treatments (e.g. pregnancy, partial gastrectomy, antibiotic treatment), which affect results from microbiological assays, but not from the radio-isotopic methods. Some early work by Britt, Bolton, Cull and Spray (1969) suggested that there was a strong correlation between results from both methods. However, Raven, Robson, Morgan and Hoffbrand (1972) showed that slight discrepancies existed between methods.

In our previous notation the influencing factors θ_{ij} for the i^{th} patient serum, are constituted from such clinical factors as pregnancy, partial gastrectomy, antibiotic treatment, folate deficiency etc. We give a brief description of both methods showing that although they reflect the vitamin B_{12} present, they do so in different ways.

The microbiological assay methods rely on the increase in turbidity in the presence of vitamin B_{12} of certain bacteria e.g. Lactobacillus leichmannii, Euglena gracilis. As we have noted before, no assay method is completely specific. In such assay methods it is reasonable to suppose that antibiotic treatment will suppress the growth and hence lead to lower vitamin B_{12} measurement.

To use these methods a 'standard B_{12} curve' is constructed using different prepared B_{12} concentrations. The specimen measurements are read from this curve.

Apart from the problem of unwanted interference from other aspects, these assays reflect the actual bacteria density. The rate and amount of growth is dependent on the B_{12} present in the patient specimen. After incubation, the measurements made reflect those bacteria able and currently dividing and those incapable of further division. Further,

the 'standard B_{12} curve' is constructed in such a way that it usually gives greater sensitivity in the normal ranges.

The radio-isotopic methods rely on the mixing of a known amount of radioactive B_{12} with patient's serum. This mixture is added to a medium containing a binding protein. The radio activity of the binding protein is measured and is inversely proportional to the vitamin B_{12} present in the patient's serum. It is likely that these methods give a linear relationship between the vitamin B_{12} in patient serum and the value reported. Further, they are not affected by the clinical factors such as indicated for the microbiological assays. Unlike microbiological assays, the radio-isotopic techniques can report reliable measurement for vitamin B_{12} concentrations down to zero.

The best and most expedient practical way of dealing with non-equivalent comparisons is simply to detail, precisely, the specimens which result in wildly different measurements as compared with other, possibly standard methods. This advice is reasonable provided we measure each specimen simultaneously by two or more methods. This obviously cannot happen. We must therefore rely on the identification of inconsistent measurements to alert our attention to the possibility of an interaction between a specific type of patient specimen and the measuring method. When comparison investigations are being made it would be expedient to remove known or suspected aberrant specimens from the comparison, thus reducing the investigation to an equivalent comparison, having all θ_{ij} identically zero.

4.9 Conclusions.

The area investigated in this chapter related to the comparison of different measuring processes for the same clinical parameter. This is a natural and a necessary extension to the discussions of chapter 3. We identified an immediate problem in these comparisons. To make a justified comparison, we must compare 'equivalent' measuring processes. These are processes which measure the same basic clinical parameter and are not seriously affected by other factors which may be present. Discussions of the comparison of equivalent processes formed the basis of the chapter, and only in the last section did we discuss the problems and possible solutions which transform the comparison of non-equivalent processes into a comparison of equivalent processes.

Having ascertained that the measuring methods are equivalent, or have been made equivalent, the comparison of the processes could be viewed in two ways. Firstly, the diagnostic information contained in the results may be compared. Basically this means the comparison of the degree to which the clinical assessment of the patient would differ by using the different methods. A number of indices were given and compared. We also suggested the use of a loss function. An index was constructed which measured the 'loss' incurred by misclassification by the different methods as compared with a standard or reference process. This index, although attractive, led to difficulties in the meaning and estimation of the loss function. All of the indices discussed suffered from the same problem, and that was that they could not possibly reflect all the information available or needed in such comparisons in a simple value.

We next considered 'classical' methods of analysis. It was shown that in only very few circumstances would the product moment correlation coefficient give reliable information. Further by examining more closely the genesis of the data for comparison, we showed that the use of classical regression analysis should be dismissed. However, with some concommitant information, e.g. knowledge of one precision or the ratio of precisions, 'orthogonal' regression would prove useful.

Following from this, we investigated some of the standard methods of analysis of data arising from such comparisons. It was found that, as in the 'orthogonal' regression situation, we required 'extra' information before estimates could be computed.

Two further solutions were investigated. The first was basically a components of variance analysis. The second was an analysis suggested by Grubbs. In the first case information was available concerning the precisions of the different processes. In the second case no 'extra' information is required since one of the underlying assumptions is that no proportional error exists between processes. This assumption prevents the problems which necessitates 'extra' information. Because of the small amount of material in each specimen, we investigated designs which were combinations of these two solutions. These would allow for efficient estimation of the parameters of interest.

We can summarise the findings of this chapter in the following way. Given a situation in which two processes are to be compared, we require to obtain information concerning the accuracy (relative if necessary) of both processes and also information concerning the

precisions of the processes. (Other comparisons may be required in certain specific cases, e.g. a comparison of carryover).

A simple experiment in which patient specimens (covering the complete clinical range) are measured by both processes should be carried out. The resulting data should be plotted on a scatter diagram. The calculation of a correlation/association index, will not usually provide any more relevant information other than can be gleaned from the scatter diagram. The scatter diagram is capable of indicating peculiarities in the relationship between the measuring processes, e.g. perhaps indicating types of patient specimens which give wildly discrepant values, or showing some non-linearity in pathological ranges.

The specific clinical situation will often determine the type of analysis which should or can be carried out. If we assume, for example, that no proportional errors are present - a Grubbs analysis is possible. However, if this is not so, we require information concerning the precision of at least one of the methods, or the ratio of the precisions. Whatever the information, we reject the use of simple linear regression. A weighted orthogonal regression analysis appears to be reasonable in most circumstances. However, as we have stated the particular design of experiment is dependent on such things as the availability and volume of blood specimens. If there are blood specimens of sufficient quantity and of sufficient number the replicate design which has been described appears to be an efficient design of experiment as well as not being so complicated that laboratory staff find difficulty in following the required procedures.

If only small volumes of patient speciments, or other blood specimens, are available, it is essential that as well as carrying

out experiments in which the same specimen is measured on each process, experiments are also performed to establish the precision of each process at clinical levels throughout the clinical range. This can be done by replicating a number of specimens on the same process. It is particularly important to obtain such estimates for values in the pathological range; we require to know if the precision of the process is the same throughout the clinical range.

The discussion of the comparison of three or more machines/methods has been discussed briefly, and the same basic points can be made as in the above. In haematology, however, we are not often interested in comparing more than two methods/machines.

Chapter 5

Intra-Laboratory Quality Control

5.1 Introduction

In Chapter 3, we discussed the evaluation of a machine (method) for measuring a specific clinical parameter. This evaluation was only partial, being an investigation into the innate characteristics of the particular machine (method). To make a comprehensive evaluation, it is necessary to compare the results produced for patient specimens with that of a standard (reference) measuring process or at least some established method(s) which are used to measure the same clinical parameter. This problem was considered in chapter 4.

In both chapters 3 and 4, we required to make the assumption that the machines and/or methods under discussion were 'in control', that is, produced accurate, reliable meaningful results for clinical use. While discussing measuring processes in chapter 2, we indicated that unless a measuring process was 'in statistical control', the resulting values were meaningless. Since a machine/method which does not produce meaningful clinical results can have no place in a clinical laboratory, we had to assume that the measuring processes under discussion were 'in control'.

In this chapter and the next, we discuss the control of measuring processes. Control of a clinical measuring process will be seen to have two components - Intra and Inter laboratory control. The first of these, Intra Laboratory quality control will be discussed

in this chapter. The second, Inter Laboratory quality control, will be discussed in chapter 6.

This chapter opens with an investigation into the meaning and use of 'normal ranges' in clinical situations. Apart from being based on the measurements from the laboratory, these ranges are frequently used in the control of measuring processes. We next discuss, in general, certain methods which have been proposed for implementation in control schemes. Finally we discuss a complete scheme of control which is currently being used in a haematology laboratory.

5.2 Normal Ranges (Reference Ranges)

5.2.1 Introduction

A comprehensive review of 'normal ranges' is given by Martin, Gudzinowcz, and Fanger (B 28). The review is mainly concerned with situations arising from Clinical Chemistry. We shall investigate the concept, meaning, construction and use to which 'normal ranges' are put.

Variation is apparent to any person interested in measuring 'normal' or indeed 'abnormal' biological parameters. (We consider any measurement to be made using a statistically controlled measuring process. This will be defined in 5.3.1). Often the expected value, average or mean, of the clinical parameter being measured becomes confused with the concept of normalcy of measurement. A single 'normal' value has no practical meaning, since due to the biological variability, very few people will commonly have the specific value. Hence a range of values is considered, such that all values in this interval are considered as 'normal'.

Before proceeding with the discussion of the usage and understanding of this concept, we discuss the variety of meanings which can be given to the term 'normal range'. Murphy (1972) has given a list of different meanings which can and have been given to the word 'normal' in the phrase 'normal range'. He arranges them in increasing order of abstraction and suggests alternative terms. The list given below was presented by Galen and Gambino (B 16).

Paraphrase	Where Used	Preferred Terms
1. Probability Function.	Statistics.	Gaussian.
2. Most Representative of its class.	Descriptive sciences.	Average, Median, Modal.
3. Commonly encountered in its class.	Descriptive sciences.	Habitual
4. Most suited to survival.	Genetics, Operation Research	Optimal, 'fittest'.
5. Carrying no penalty.	Clinical Chemistry	Innocuous, harmless.
6. Commonly aspired to	Politics, Sociology	Conventional
7. Most perfect of its class.	Metaphysics, Ethics, Morals.	Ideal.

The use of the word in 1. has little to do with the sense in which it is used in any of the other cases. Galen and Gambino point out that Gauss originally introduced the Normal Curve to describe the distribution of errors experienced when repeat measurements were made on the same object. Now, although this is strictly correct, it is too restrictive today. The importance of the Normal Curve other than its use when considering replicate measurements is well documented. It is generally accepted that many clinical parameters have distributions

which can be represented by this bell-shaped Normal, or as we have called it, Gaussian curve.

Murphy (1972) suggests that the laboratory worker defines the 'normal range' as the most commonly encountered values, i.e. as in 3. The clinician, on the other hand, deals with a single patient, not caring about the abstract statistical concepts, only wishing to know whether the result obtained is threatening to life, (or is an indication of some pathology). Galen and Gambino suggest that the clinician's concept is that of 4. However, I think that this is a simplification. In some situations this will be true, but in other cases the clinician will quite probably understand the concept as harmless as in 5.

There is no doubt that there is confusion over the meaning and usage of the phrase 'normal range'. We shall see that in most cases, the meaning which presumably underlies most of the current thinking, is that which is given in 3, i.e. most commonly encountered in its class.

We have noted the use of 'normal' in the statistical sense. That is its use in describing the familiar bell-shaped curve which we have called Gaussian. From Folin and Berglun (1922), histograms and frequency curves have been used to represent pictorially the distribution of the measured values of clinical parameters. With such pictorial representations, came an awareness that the distribution of measured values could be symmetric or skewed. Frequently, if skewness is present, it will be such as to produce a 'tail' to the right of the central maximum i.e. positive skewness. We shall show that if this is so, that the use of a 'normal range', $\bar{x} \pm 2s$, will have few of the properties required of the meanings given by Murphy.

When constructing a 'normal range', we must decide exactly what class of people are being described. Fundamental to the investigation is the concept of the statistical population to which we are referring. This point is emphasisied by Healy (1969), who suggests that in many instants, the definition of the population may not be as easy as at first imagined.

Let us suppose that we wish to measure some clinical parameter which shall be denoted X. Let us present three populations relating to this parameter.

(a). The set of all values of X, which people, of a specific type and at a specific time, exhibit. (Type could define a group by age, location etc.)

(b). The set of all values obtained by a specific measuring process for the population defined by (a).

(c). The set of all values obtained by a specific measuring process for a definable and attainable subset of the population defined in (a).

The population defined by (a), is theoretical and can never be observed. However, in a sense it is the 'true population'. The observed population defined by (b), is not equivalent to that in (a). We obviously cannot measure all persons at the same time and further, errors are introduced due to the particular measuring process (e.g. machine error, drifting of the process, sampling of the blood specimen, reagents used in determination etc.) The population defined by (c) is somewhat more practical - it defines an attainable group of people. Usually there is an implicit assumption that (c) is a random sample from (b). The populations (a) and (b) are easily defined, but are not realistic - they cannot be investigated.

To obtain 'normal ranges', it is obvious that we require to deal with 'normal', i.e. apparently healthy persons. This, an easy defining relation, is not as simple to use in practice as first appears. Obviously, we cannot measure the clinical parameter even for the total population, (c) of healthy people. Hence we resort to sampling from this population.

If the clinical parameter X, has the property that an observable pathology occurs when certain values of X are observed, then the decision 'normal'/'abnormal' is predefined and to a certain extent obvious. (We note that investigations have been undertaken to study differences between visual observations made by clinicians.)

However, if as usual the measurement of X provides a 'yardstick' for assessing 'normality', we see that the decision relating to whether a person is 'normal'/'abnormal' i.e. whether we include or exclude a person in the 'normal' population, is confounded with the construction of the 'yardstick'. Thus sampling from population (c) contains an element of subjectivity due to the defining of the population.

Just as there is innate variability in 'normal' measurements, variability occurs in the 'abnormal' measurements. We are confronted with the classic problem of overlapping populations. Basically we are required to judge which of the populations a specific specimen (person) comes from. We have the usual classification of false positives and false negatives, see section 4.2.3.

The situation is rendered even more difficult because it is known that the parameter X, could conceivably be affected by one or more of the following factors:-

age, sex, diet, exercise, emotion (stress), posture, tourniquet techniques, drugs, fluid intake, racial and genetic factors,

time of day, month or year, menstrual status etc.

(A further problem, discussed later, is that abnormality in one clinical parameter does not necessarily imply abnormality of another parameter.)

In the definition of population (c), a specific (usually the standard or reference) measuring process is implied. We have discussed in Chapter 4 the comparison of different measuring processes. Further in section 4.8 we indicated that some different processes cannot be regarded as being equivalent. The specific example given was the determination of serum vitamin B_{12} by a microbiological method and by a radio-isotopic method. If, for example, a patient has a 'normal' serum vitamin B_{12} level and has recently been medicated with an antibiotic drug, it is possible that the radio-isotopic method produces a 'normal' result while the microbiological growth method produces an 'abnormal' result, (decreased). This shows that there is a difficulty in defining the populations (b) and (c).

It will come as no surprise to find that 'normal ranges' constructed in different regions (locations) often differ.

Let us now consider the construction of a 'normal range'. We discuss two approaches - Parametric and Non-parametric.

5.2.2 Construction of Normal Ranges: Parametric

Such an approach depends on the assumption of a mathematical formulation for the underlying distribution of the measured clinical parameter, e.g. Gaussian, log-Gaussian etc. (We note that this assumption is more demanding than, for example, the assumption of a continuous measurement scale).

A common 'normal range', which has been suggested by many authors, is

$$(\mu - 2\sigma , \mu + 2\sigma)$$

where μ represents the population mean and σ represents the population standard deviation. That is the 'normal range' consists of all values within the interval - plus/minus two standard deviations from the population mean. However, this cannot be of practical use, since neither μ nor σ will be known. In these situations μ and σ^2 are replaced by the unbiased estimates $\bar{x}$ and s^2 respectively. We note that although s^2 is an unbiased estimate of σ^2 it is not true that s is an unbiased estimate of σ. We consider using the interval

$$(\bar{x} - 2s, \bar{x} + 2s)$$

This is a random interval, dependent on the sample from which it was constructed. Not only will the location of the interval change from sample to sample, through $\bar{x}$, but the size of the interval will also change, through s.

The interval $\bar{x} \pm 2s$ has been adopted by the American Society of Clinical Pathologists as the definition of the 'normal range'. We remark that some caution is required. If the distribution of the measurements made on the clinical parameter has a Gaussian distribution, then the interval $\mu \pm 2\ \sigma$ contains 95.45% of the population. The interval $\bar{x} \pm 2s$ does not have this property, even in the 'long run'. The expected proportion contained by this interval depends on the sample size. An interval of the form $\bar{x} \pm \lambda s$ can be constructed to give an expected proportion of γ in the long run. This interval is:-

$$\bar{x} \pm t_{n-1,\gamma} \sqrt{\frac{n+1}{n}}\, s$$

where $t_{n-1,\gamma}$ is the γ percentage point of the student t-distribution, with $n-1$ degrees of freedom.

Intervals of the form, $\bar{x} \pm \lambda s$ whatever λ may be, are probably only of use if the underlying distribution is Gaussian. Even with symmetric distributions, the probability statements which we have made are not valid. The reason for the difference is based on the fact that $\bar{x}$ and s^2 are jointly sufficient statistics for the Gaussian mean and variance, i.e. location and spread. Hence for the Gaussian distribution all the information relating to position and scale is contained in $\bar{x}$ and s^2 (or in μ and σ^2). (A sufficient statistic for a statistical parameter corresponds to the greatest reduction of the data without loss of information relating to the parameter.) This is not true for distributions other than the Gaussian.

When the underlying distribution is not Gaussian but is symmetric, the interval $\bar{x} \pm \lambda s$ will usually produce acceptable ranges, which do not carry the properties which we have stated above. If the underlying distribution is skewed, the interval or some part of it may contain some meaningless values. Consider the following example.

Suppose that the underlying distribution of measurements, X, is log-Gaussian, (i.e. log X has a Gaussian distribution). Further, let log X have mean 0 and variance 1. (This means that X has mean $e^{\frac{1}{2}} \approx 1.65$, variance $e(e-1) \approx 4.67$ and skewness $(e+2)\sqrt{(e-1)} \approx 6.185$.) We wish to calculate the 'normal range' as described above on the basis of a sample drawn from the population. The sample given in table 5.1 below has been generated from such a population.

TABLE 5.1 on next page

TABLE 5.1

0.5904	2.0340	0.4387	1.8851	1.4978	4.6927	0.4202	1.4667	0.4140	0.7619
0.3746	0.4858	1.0387	0.3926	10.6547	1.1196	0.1456	1.1901	0.7741	0.4790
1.4506	2.1426	1.8739	0.2236	1.6854	0.9296	0.7386	5.3388	8.3311	2.8548
2.0585	1.2712	1.6274	0.9484	0.5798	0.7468	4.2207	0.8065	5.9715	3.9432
3.3906	2.1728	0.3227	0.2567	1.0335	1.6938	0.2979	0.5174	7.3522	1.1960

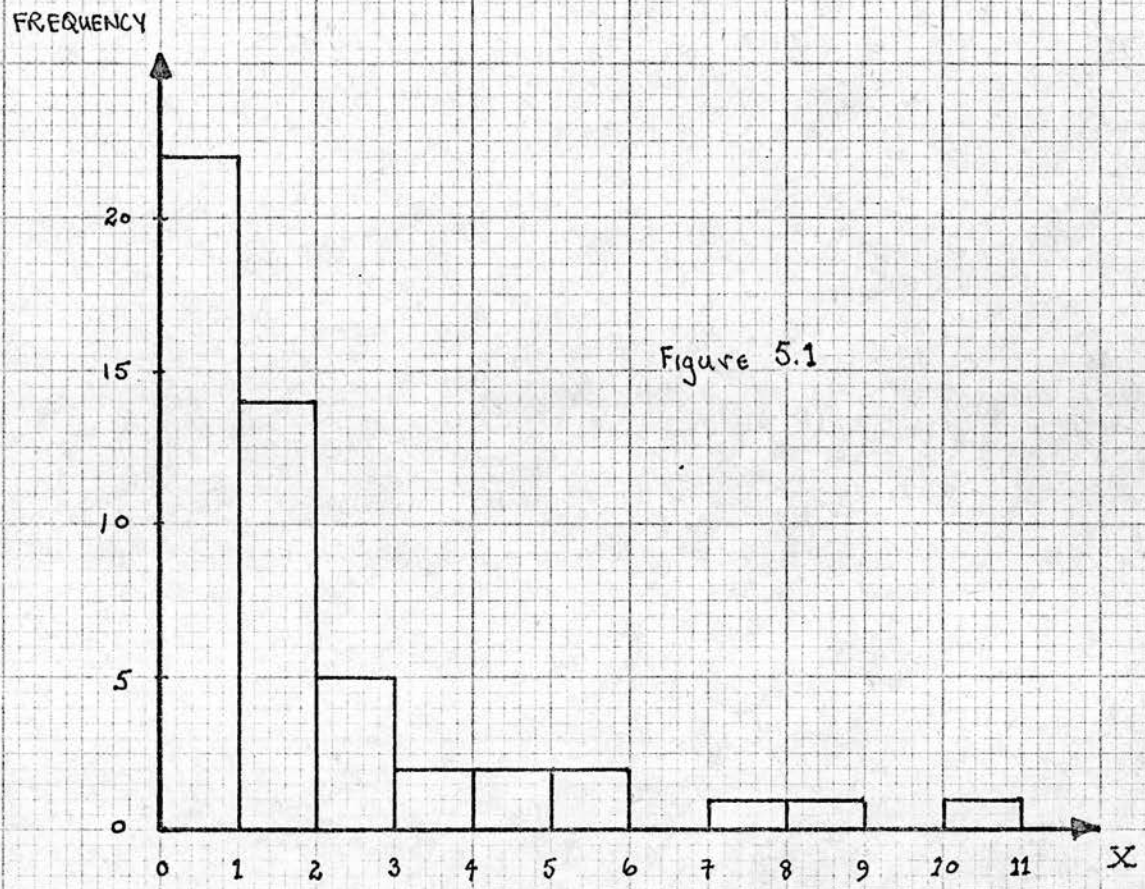


Figure 5.1

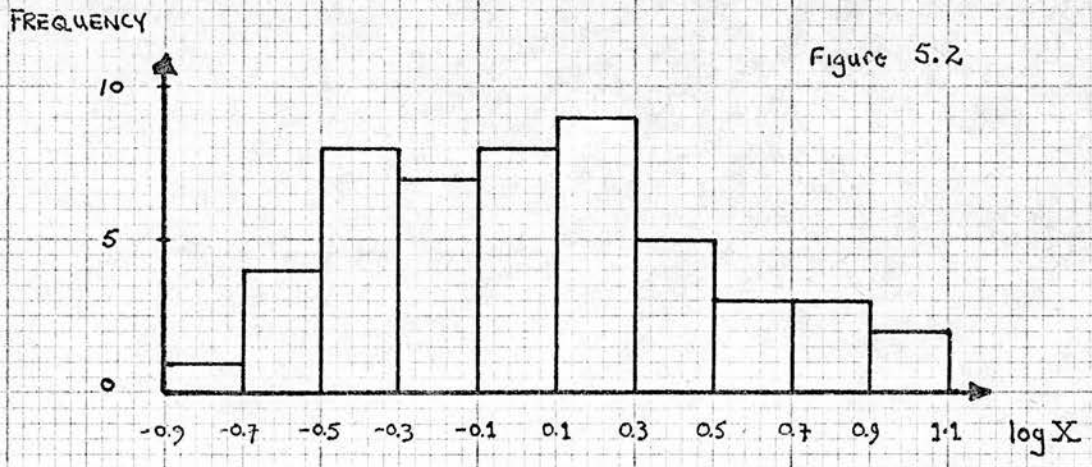


Figure 5.2

WELL... A4 210 x 297 mm

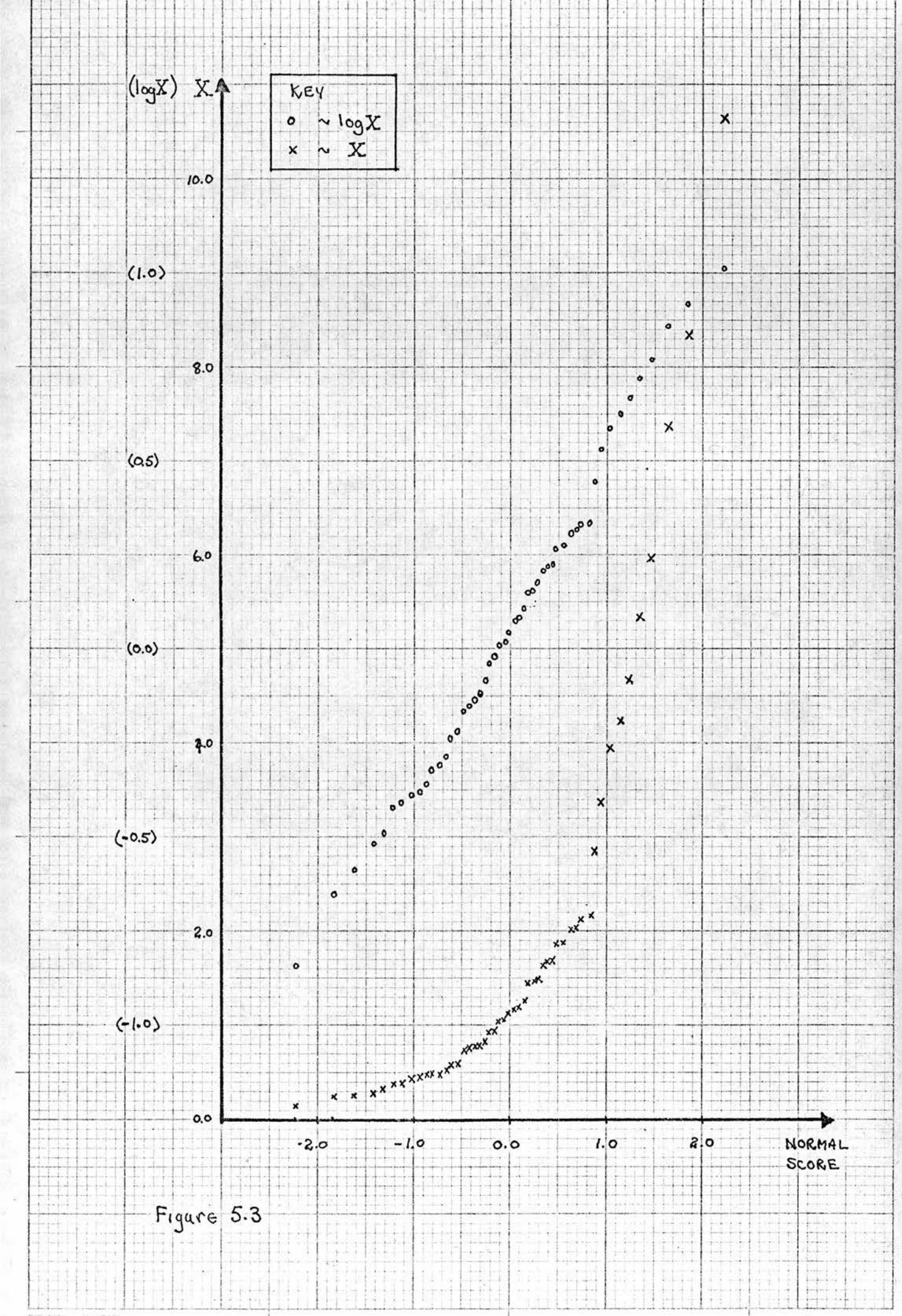


Figure 5.3

WELL
A4 210 x 297 mm

Figure 5.1 shows the corresponding histogram for these data. We note that the histogram shows a marked skewness. If we construct the interval, $\bar{x} \pm 2s$ we obtain (-2.5207, 6.3941). More than one quarter of the interval covers impossible values for the 'clinical measurement', i.e. negative values. If we were to transform the data, dealing with the logarithm of the measurements, the resulting histogram would be as shown in figure 5.2. The histogram has a much more acceptable symmetry. In itself this would be a reasonable justification for the use of the interval, $\bar{x} \pm 2s$. However, we could also plot the graphs given in figure 5.3. These graphs show the ordered data set and the ordered logarithmic data set plotted against the expected order statistic of the Gaussian distribution. Tables of these 'Normal Scores' are given by Fisher and Yates (B 14) and Neave (B 29). (Probability paper may be used in place of these tabulated values.) The graph of the plotted points is approximately straight when the data is Gaussian. It will be seen that the transformed data is approximately linear, unlike the original data. The interval corresponding to $\overline{\log x} \pm 2s_{\log x}$ is (-0.8045, 0.9394). Exponentiating this, we obtain (0.1569, 8.6975). From the histogram of the original data, figure 5.1, we see that this interval is more reasonable, than the first. Of course there is no suggestion that this interval has any greater meaning clinically.

In the example which we have given, we knew which transformation to use to give a reasonable fit to a Gaussian distribution. However, in practice we would have to investigate a number of transformations, if the raw data was skewed. Henry (1960) has discussed this problem in more detail.

The interval, $\bar{x} \pm 2s$ is, of course, arbitrary. Mainland (1971) remarks that such an interval is not in his considered opinion a good biological standard of 'rightness of health'. Wootten and King (1953) working with non-symmetric distributions for blood constituents used two ranges - an inner range containing 80% of normal findings and an outer range containing 98%. Measurements lying within the inner interval were considered unquestionably 'normal' with those outside the outer interval unquestionably 'abnormal'. Those measurements not classified by these means were considered questionable (see later). This type of procedure can be used for any distribution, not necessarily giving symmetric ranges about the mean, median or mode. Mainland (B 26), makes no definitive statement concerning the 'normal' range, but tabulates the percentage of the population excluded by $\bar{x} \pm ks$ for various values of k. We note, once more, that such statements depend on n, the sample size.

Tolerance intervals, a concept due to Shewhart (1931), have been used by Henry (B 19). The basis of a tolerance interval can be described in the following way. On the basis of a sample of n independent observations, an interval (L_1, L_2) is constructed for which some probabalistic statement concerning a given proportion of the population γ can be made. These statements can be made in one of two ways. Firstly, probabalistically - a tolerance interval (Atchison and Dunsmore (B 1) refer to this as a guaranteed tolerance interval) such that with given probability β at least a proportion γ of the population lie between L_1 and L_2. Secondly, we could construct what Atchison and Dunsmore call expected mean tolerance limits - an interval which covers a given proportion of the population on the average.

For the Gaussian distribution, the problem of a guaranteed tolerance interval has an approximate solution. As we have seen above, the construction of intervals are based on $\bar{x}$ and s^2. (Since they are sufficient statistics for μ and σ^2. Further, intervals of this form are invariant under transformations f: $\mu \to \mu+a$ and $\sigma \to b\sigma$).

Let $\delta(\bar{x}, s^2, \lambda) = (\bar{x} - \lambda s, \bar{x} + \lambda s)$

and $g(\delta, c) = P[P(\delta(\bar{x}, s^2, \lambda)) \geq c]$

Without loss of generality, we can consider that the underlying distribution has mean 0 and variance 1.

Let $f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \quad -\infty < t < \infty$

Then $P[\delta(\bar{x}, s^2, \lambda)] = \int_{\bar{x} - \lambda s}^{\bar{x} + \lambda s} f(t)dt$

We seek λ, so that

$g(\delta, c) = \beta$

i.e. $P[P(\delta(\bar{x}, s^2, \lambda) \geq c] = \beta$

Kendall and Stuart (B) give a method for obtaining an approximation using the conditional probability

$P[P(\delta(\bar{x}, s^2, \lambda) \geq c | \bar{x}]$

We use the following approximation:-

$$P[P(\delta(\bar{x}, s^2, \lambda)) \geq c | \bar{x}] = P\left[\chi^2_{n-1} > (n-1)\frac{r^2}{\lambda^2}\right]$$

where $\int_{\bar{x}-r}^{\bar{x}+r} f(t)dt = c$

$$\text{and } P[P(\delta(\bar{x}, s^2, \lambda)) \geq c] = P\left[P(\delta(\bar{x}, s^2, \lambda)) \geq c \,\middle|\, \bar{x} = \frac{1}{\sqrt{n}}\right] + O(n^{-2})$$

is used to obtain an approximate value for λ.

Howe (1969) gives a slightly different procedure for approximating λ and also obtaining the order of the associated error.

Mean tolerance intervals can be constructed as follows. Consider the interval given by:-

$$\delta_*(\bar{x},s^2,\lambda) = (\bar{x} - \lambda s, \bar{x} + \lambda s)$$

We require the following condition to hold:-

$$E\left[P(\delta_*(\bar{x},s^2,\lambda))\right] = c$$

If we are dealing with an underlying distribution which is Gaussian with zero mean and unit variance (no loss to generality), then we know that the distribution of the sample mean is Gaussian with mean zero and variance $1/n$, where n is the size of random sample taken. Further, we know that $(n-1)s^2$ has a Chi-square distribution with $n-1$ degrees of freedom. We require to solve the following equation for λ.

$$\int_{-\infty}^{\infty} \int_{0}^{\infty} \left[\Phi(\bar{x}+\lambda s) - \Phi(\bar{x}-\lambda s)\right] f(\bar{x})g(s)d\bar{x}ds = c$$

Where $\Phi(u)$ is the cumulative distribution function of the Standard Gaussian function. On solution of this we obtain

$$\delta_*(\bar{x},s^2,\lambda) = \left(\bar{x} - s\sqrt{\frac{n+1}{n}}\, t_{\frac{1+c}{2}} \,,\ \bar{x} + s\sqrt{\frac{n+1}{n}}\, t_{\frac{1+c}{2}}\right)$$

where t_α is the $\alpha\%$ point on the Student-t distribution.
Both types, guaranteed and mean tolerance intervals are tabulated in Geigy, (7th edition) Scientific Tables (B 18).

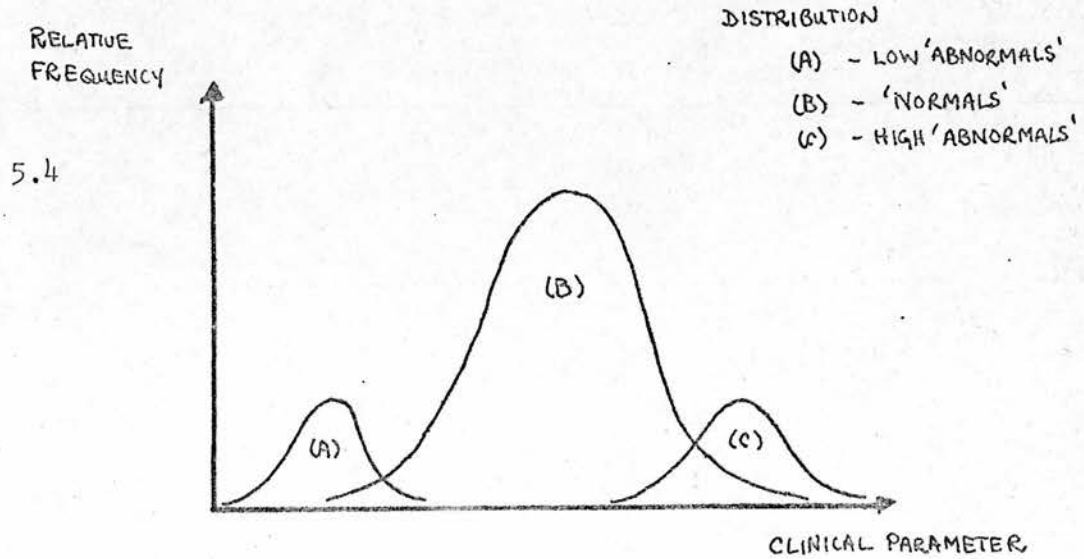
The tolerance intervals which we have described are symmetric about $\bar{x}$. This implies that 'abnormality' is associated with values at either end of the clinical scale. Furthermore, the 'abnormal' classification is assumed to occur at the same distance from the population mean. There are situations, in practice, in which this

is not appropriate. Either we require asymmetric tolerance intervals, or we require one-sided tolerance intervals. In the first case, the percentage 'abnormal in each tail of the distribution must be known. Thus we allow for situations in which 'abnormality' occurs at different distances from the mean (average) clinical parameter value. In the second case the percentage 'abnormal' and the tail in which these occur must be known. Thus we allow for the situation in which no 'abnormality' is associated with a particular extreme of clinical values. In both situations, the methods of construction are slight modifications of the methods which have been given for the two-sided symmetric case above.

The problems encountered in the above, assuming that the underlying clinical distribution was Gaussian, are usually multiplied if some other distributional form is assumed.

Another situation may be assumed to exist. In this situation, the total biological population is assumed to consist of sub-populations. For example, assume that the 'normal' sub-population has a Gaussian distribution. Let us assume that the 'abnormal' sub-populations, high and low, also have Gaussian distributions. Schematically we have the situation shown in figure 5.4.

FIGURE 5.4



Standard methods are available for facilitating such decompositions - using probability paper or non-linear least squares curve fitting programs. Such decompositions are only necessary if we believe that the population under investigation is heterogeneous i.e. contains 'normals' and 'abnormals' such that a single mathematically tractable distribution cannot describe accurately the situation. In many situations authors tend to assume a distribution for the 'normals' and consider the 'abnormals' to be outliers with respect to this distribution.

As an example of problems arising from such decompositions see Prescott, Hedley and Paterson (1973). They indicated that the decomposition into 'abnormal' and 'normal' with respect to serum alkaline phosphatase, which was presented by Cooke, Swan, Asquith, Melikian and McFeely (1973) was sensitive to the assumption that the 'normals' had a Gaussian distribution. (This assumption was not supported by data presented, nor by other clinical evidence). The outcome of the original study suggests that there is widespread vitamin D deficiency in the population. However, as the criticism points out this is not supported by a more appropriate analysis of the data. Other comments made by Prescott, Hedley and Paterson underline points which have been made in our discussion.

5.2.3 Construction of Normal Ranges : Non-parametric

In this section we consider the construction of 'normal ranges' making no assumptions, except perhaps one of continuity, concerning the distribution of the clinical parameter in which interest is centred. In the previous section we saw that two types of tolerance intervals could be used as a 'normal range'; guaranteed tolerance interval and mean coverage tolerance interval. The same choice is open

when we consider non-parametric construction of tolerance intervals.

Non-parametric mean coverage tolerance intervals are based on percentiles (or some other n-tiles) of the collected data. To estimate the interval (L_1, L_2) for which $100\gamma\%$ of the population are contained on average, we require to know the proportion of the population below L_1 or the proportion above L_2, $(\gamma_1 + \gamma_2 = \gamma)$. The points L_1 and L_2 are estimated by the 100γ, percentile and the $100\ (1 - \gamma_2)$ percentile of the data. As before, the estimation of the tolerance interval is dependent on the amount of data available. Basically, for such an estimation to be reasonable a large data set is required.

We have previously noted that Wotten and King (1953) have used this type of approach, and others have followed e.g. Herrera (1958). The use of percentiles can be made to great effect when the underlying distribution is non-symmetric.

Guaranteed tolerance intervals can also be constructed. Wilks (1941, 1942) showed that an interval (L_1, L_2) can be constructed so that

$$P\Big[\int_{L_1}^{L_2} f(x)dx \geq \gamma\Big] = \beta$$

where $f(x)$ is the unknown continuous density function of the clinical parameter measurements. The interval (L_1, L_2) is based on order-statistics. Robbins (1944) showed that only order-statistics can solve this problem. Discrete distributions can be similarly handled although some minor alterations are necessary.

Let our sample of measurements be $(x_1, x_2, \ldots, x_n)$ and the ordered sample, in ascending order of magnitude be $(z_1, z_2, \ldots, z_n)$ so that $z_1 \leq z_2 \leq \ldots\ldots z_{n-1} \leq z_n$.

Let $L_1 = z_L$ and $L_2 = z_U$, and let $F(x)$ be the cummulative distribution function of the measured clinical parameter. Thus we can write

$$P\left[\int_{L_1}^{L_2} f(x)dx > \gamma\right] = P\left[F(z_U) - F(z_L) > \gamma\right] = \beta$$

Now we see that $y = F(z_L) - F(z_U)$ is a random variable whose distribution can be constructed and can be shown to have a Beta form. Thus we can write

$$P\left[y > \gamma\right] = \int_{\gamma}^{1} \frac{z^{U-L-1}(1-z)^{n-U+L}}{B(U-1,\ n+L-U+1)}\ dz = \beta$$

For given γ and β, and assessing symmetric intervals this can be solved for either U or L by using Incomplete Beta Tables.

Example: 'Normal Ranges'.

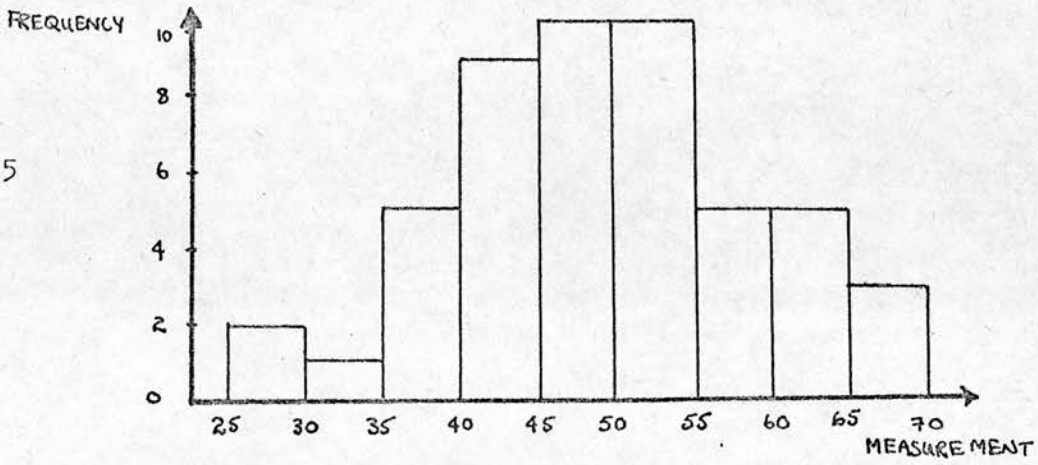
To illustrate the above work, we take an example and construct the ranges which we have discussed above. The data, fifty 'measurements', are a random sample from a Gaussian distribution with mean 50 and variance 100. We have taken such an example (see table 5.2) in order to compare our results with the theoretical range. Ordinarily, we would require to take more than fifty measurements to obtain clinically justifiable ranges, but we do not wish to make our example larger than required to pick out the essential features of the above discussion.

TABLE 5.2

48.11	47.60	51.33	27.10	43.84	45.63	54.59	45.01	58.45	53.83
68.66	36.02	50.68	50.53	29.66	64.26	62.54	60.67	55.92	51.74
49.82	56.28	52.30	56.59	47.02	69.27	47.18	57.69	43.10	66.75
43.54	46.50	53.24	33.25	61.90	39.24	62.87	35.74	53.45	47.85
38.50	47.80	44.67	59.12	42.90	40.96	41.83	38.40	40.81	43.41

Although we know that this data comes from a Gaussian distribution, we shall construct the associated histogram to check symmetry, at least visually, see figure 5.5

FIGURE 5.5



From the data we compute the following statistics

$\Sigma x = 2468.15$ $\Sigma x^2 = 126571.8101$ $n = 50$

$\bar{x} = 49.36$ $s^2 = 96.66$ $s = 9.83$

Since we know the exact population from which the sample was taken, we can compute the 'normal range', $\mu \pm 2\ \sigma$. In usual practice this cannot be done. In table 5.3, we present the numerical values for the 'normal ranges' which we have discussed.

TABLE 5.3

Normal Range	Numerical Value
Parametric.	
$\mu + 2\sigma$	30 - 70
$\bar{x} + 2s$	29.70 - 69.03
Guaranteed	25.97 - 72.75
Mean Coverage	29.41 - 69.31
Non Parametric	
$2\frac{1}{2}$ - $97\frac{1}{2}$ percentiles	28.12 - 67.91
Order Statistic	29.66 - 68.66 ($\beta \approx 0.06$)

5.2.4 Discussion

After 'normal' ranges have been constructed, they can be used in a number of different ways, which lead Martin, Gudzinowicz and Fanger (B 28) to say:-

'However, the statistically defined 'normal' range (to encompass all values within the limits set by the mean $\pm$ 2 σ), until its formal adoption in 1968 by the American Society of Clinical Pathologists, has been interpreted in the literature in various ways by different scientists and authors, and its true meaning today is still in doubt'.

The 'normal' range is used as a diagnostic tool, as a support for a diagnosis and as an aid in the management of disease. However, it is doubtful whether it can convincingly manage any of these in much more than a rough manner. The interval, be it $\bar{x}$ + 2s or (z_L, z_U) suffers the same criticisms e.g. underlying assumptions as the subjectiveness of such an index as a measure of 'rightness of health'.

Essentially, we wish to construct a predictor $p(\underset{\sim}{x}_i | \underset{\sim}{y})$, where $\underset{\sim}{x}_i$ is an information vector for the i^{th} patient and $\underset{\sim}{y}$ is an information matrix based on past clinical experience. The predictor gives a probability (or other measure) of the i^{th} patient being 'ill'. No specific disease is under investigation. To some extent this is 'wishing for the moon' and perhaps if we reduce the problem,which has been done in some cases, to $P_d(\underset{\sim}{x}_i | \underset{\sim}{y})$ where the predictor gives likelihood of certain specific diseases, we may stand a better chance of succeeding. This constructs a predictor of specific diseases, not of general health.

With the increase in the use of machines such as the Coulter-S and the Technician SMAC, many clinical parameters can be determined

simultaneously. Best, Mason, Barrow and Shepherd (1969) have noted:-

'The question of a 'normal profile' in multiphasic screening has not been adequately aired from the probability point of view. Concern has been expressed because of the great number of 'normals' who show one or more 'abnormalities' in a screening examination. It is apparent from our computations and considerations that if sufficient tests are made in a truly 'normal' population, it will be the rare person who has a completely 'normal profile'. This is a matter of mathematics, not public health.'

Schoen and Brooks (1970) suggest that the 'normal range' for each patient be redefined as a function of the number of tests in the multiphasic screening - in general for n independent trials with overall probability level of 95%, a single test probability level, p, is obtained from

$$0.95 = p^n$$

Schoen and Brooks further suggest that a patient's results outside $\bar{x} \pm 2s$ but within $\bar{x} \pm 3s$ may be followed up, provided that this is suggested by other information. While if results are outside $\bar{x} \pm 3s$, they should be followed up as a consequence of a known abnormality.

One of the assumptions, of course, is that n independent trials are made. This is obviously not often true - measurements are correlated - and this affects the setting of individual probability levels.

When defining the populations (a), (b) and (c) earlier, we omitted a different type of population, which may be considered.

(d). The set of all possible measurements by a specific method which could be made on a single person.

Investigations of such a population would enable us to establish personal 'normal ranges'. Such an idea is not new, Rietz and Mitchell(1910)

investigated and questioned the usefulness of such a procedure. Obviously, the stability of measurements on an individual depends on the clinical parameter being measured. Basically, the same problem occurs; when do we consider a measurement to indicate abnormality?

5.3 Monitoring a Laboratory's Clinical Measurements

5.3.1 Introduction

In this section we shall discuss intra-laboratory quality control, that is to say the quality control carried out by the individual laboratory. Essentially this control compares the laboratory's current state of control with its own past results. In other fields of clinical investigation, this type of quality control also includes the comparison of the laboratory's results with standard preparations, etc., however in the majority of cases in haematology this is not the case. There are few standards against which to compare.

The International Committee for Standardisation in Haematology has investigated the particular problems involved in establishing reference and standard preparations for haematological determinations. Some success has been achieved. For example, a reference preparation for photometric haemoglobinometry is available. A solution of cyanmethaemoglobin (HiCN) is used, the specification being given in British Standards BS 3985: 1966.

We shall discuss some of the most common methods of controlling the measuring process. These techniques can be used in other situations, but we shall explain them with special reference to intra-laboratory quality control These control methods can be used for Primary or Secondary measurements.

Hoffmann(B 21) in his book 'Establishing Quality Control and Normal Ranges in the Clinical Laboratory' has likened quality control in a laboratory to the practice of medicine.

'Real quality control is similar to the practice of medicine. The good physician examines his patient, then evaluates the information gathered during the examination, making use of all his medical training and experience. At that point he will make his decision about diagnosis and treatment.

Every day, the supervising technologist should examine his laboratory. In a fashion somewhat similar to that of a physician, the chief technologist 'diagnoses and treats' his laboratory'.

Why does the technologist 'diagnose and treat his laboratory'? Essentially, he is supplying information to the clinician. The clinician uses this information as a basis for decisions concerning the patient. The more reliable the information, the 'easier' the clinician's task, and further, the better the quality of care that the patient receives.

When the laboratory's measurements are reliable, the greater is the possibility of detecting occult disease. Hence the quicker the appropriate treatment can be undertaken. With less reliable results, overt disease will probably be detected, but the possibility of early detection of occult disease is reduced. I am not advocating diagnosis by numbers, but obviously the number that the clinician calls for must correlate with other observations he makes.

A major point made in the above quotation, is the suggestion that the technologist should use all information gathered. We shall point out later that this is not always the case in practice and improvements could be made if this suggestion was implemented.

We return to a point made earlier, that for any measurement to be used in a clinical situation it must be reliable. This implies that the measurement must come from a process which is in 'statistical control'.

The measurement process must be stable or have a certain consistency. This will be reflected in the reproducibility of the process i.e. the precision of the process. Obviously the concept of accuracy is related to this. Eisenhart (1963) has stated:

'Consistency or statistical stability of a very special kind is required: to qualify as a measurement process a measurement operation must have attained what is known in industrial quality control language as a state of statistical control. Until a measurement operation has been 'debugged' to the extent that it has attained a state of statistical control, it cannot be regarded in any logical sense as measuring anything at all. And when it has attained a state of statistical control there may still remain the question of whether it is faithful to the method of measurement of which it is intended to be a realisation'.

We should be clear that if we are monitoring the reliability of our measurement process, this does not merely relate to the reliability of the device or method we are measuring by. The reliability of information is dependent on the collection of specimens, transportation, use and production of preservatives and reagents, handling of specimens within the laboratory, administrative procedures and all other factors - personal, physical or invironmental - which affect the measurement process (see Chapter 2). Usually, and in what we describe below, the main emphasis is on the actual measuring device. However,

it is important to realise that the overall reliability of our clinical information is most influenced by the least reliable of all the sub-parts of the measurement process.

We must be aware of the medical implications of quality control schemes. Whitby (1969) has stated that in the simplest terms,

'the medical objective of quality control programmes in clinical biochemistry must be to improve standards of patient care. Unreliable laboratory results can lead to wrong decisions about the management of patients, sometimes leading to fatal results. Doctors, once alerted to the possibility of unreliability in a laboratory's work, may choose thereafter to ignore chemical results whenever they fail to fit in with clinical assessments of the condition of the patients even though decisions to disregard results of chemical investigations may also have fatal consequences.

Thus at least we have the opportunity - by means of quality control programmes - to define the degree of confidence that medical staff can place upon the results of our observations'.

This quote should alert us to the fact that while trying to obtain as reliable a measuring process as possible, we must neither alienate ourselves from the real medical situation nor spend so much effort and time in attaining it that we cannot help the clinician and decision maker.

5.3.2 Standards and Reference Materials

In many instances the problem of controlling a measurement process is relatively simple. If, for example, we wish to control the lengths of manufactured steel rods, the characteristic - length - in which we are interested, is physically observable. Experienced operators may even be able to judge the lengths quite accurately and

so gross inaccuracies of production can be identified.

In haematology or clinical chemistry laboratories the clinical characteristics being measured are usually not physically observable. Further, in a certain sense, the characteristics are defined by the process which measures them. We discussed this in chapter 4. In the situation in which we measure steel rods, international standards - be it the metre or foot - are available. Further, international standardising procedures are well known and are internationally accepted. In the haematology and clinical chemistry situations we have possibly more problems.

Materials which are used in the standardisation of machines/methods are commonly called standards and/or reference materials. Lewis (1975) has suggested, and collected, the following definitions

Primary Analytical Standard - a pure chemical substance used for the purpose of assaying, or preparing a solution of unknown (known) concentration (Radin (1967)).

International Standard - a preparation of precisely defined content, with long term stability - to which an international unit has been assigned after international collaborative study.

International Reference Preparation - an international standard but without the long term stability.

National standards or reference preparations should be identical with the international counterparts. The national standard has been called a primary standard.

A reference method for analysing a particular clinical parameter is sometimes too costly or complex. A simpler routine method may well be carried out - often an automatic method. Comparisons between autoanalysers and reference methods are essential to check

for both accuracy and precision of the autoanalyser.

Practically, in a routine laboratory quality control scheme, many different standards and/or reference preparations may be used. The problem of finding an acceptable standard or reference material can be very difficult. For example, the search for an adequate standard for red cell counting is on going. Two types of materials have been considered - artificial and natural blood, preserved or modified by a fixative treatment. No artificial materials -pollens, spores, yeast, latex or other plastic polymers - have proved suitable. The lack of stability of the natural materials is obviously a serious handicap. Further, there appears to be an interaction between measuring methods/machines and preservatives. As Lewis (1975) remarks in this situation,

'a major problem with all such preparations is that their deterioration on storage results in a greater variability than the precision of the instruments which they are intended to control'.

5.3.3 Quality Control and Frequency of Monitoring

One of the first questions to be answered is, 'How often should Quality Control be carried out?' However, before we can answer this question we must examine some basic points. The first of these is, 'why is quality control being carried out?'

Basically, as we have indicated, we carry out quality control because the measurements can only have meaning if they are produced by a statistically controlled measuring process. Further, high reliability of clinical measurements allows, potentially, for the early identification of occult disease, better management of the patient etc.

The second point to examine is, what do we mean by carrying out quality control? Plainly, this is monitoring the results being produced by the measuring process. That is, performing certain quality control procedures and acting on the results from these procedures. The monitoring should cover all parts of the measuring process, which have been identified in chapter 3.

We shall describe some procedures commonly used in quality control in the following sections. However, before we return to the first question, we should identify the factors which affect the degree of control and the frequency of control which should be applied. Obviously, one of the factors is the degree of reliability which is required by the clinical staff. The accuracy and precision of the results will depend on the type of establishment - research/routine laboratory. Further, it could depend on the patient. For example, if a male patient's specimen has a reported haemoglobin concentration of 16.0 g/dl, a possible error of $\pm$ 1 g/dl would not affect the decision that the result is within the 'normal range' (16.0 $\pm$ 2.0 g/dl). However, if the reported value has been 14 g/dl with the same error $\pm$ 1 g/dl this may, depending on the specific error, place the result in the 'normal' or the 'abnormal' range. Thus, although for the majority of results which are 'well within the 'normal range'' the quality control may be thought of as 'wasted effort', we cannot know which specimens require high quality control and those which do not require such high quality. The response is to maintain a reasonable degree of quality without imposing such heavy demands on the laboratory that it cannot operate.

This brings us naturally to the effects on the laboratory's workload. These effects are variable. Unfortunately, under extremely

heavy workloads, when the measuring process is probably more prone to unreliability, the pressure is to produce results and often quality control is reduced. The last factor is achieved at the cost of the quality control. These monitoring procedures involve the use of the measuring process, which nearly always means an 'increased' workload. Further, 'extra' reagents, donor-blood, commercial reference materials, the laboratory's own reference materials etc. could be required. This will almost certainly increase the cost. However, as before we could give cost a meaning wider than its common financial sense.

In the discussion above we have used inverted commas to mark the emotive words; increase, extra. We must be clear that quality control is not extra work, and does not really increase workload. It is an integral part of the measuring process. In fact without quality control we do not have a meaningful process.

The original question remains, 'how often should Quality Control be carried out?'. Before we finally attempt to answer this question, it should be emphasised that there is no single quality control scheme which can be used in all laboratories. We have already seen that quality control should be geared to the requirements of the clinical and laboratory staff. Further, we must point out that the size of the laboratory affects any scheme for monitoring the process. That is, the size with respect to both the number of measurements requested and the number of staff. Now, as we shall see in the following sections, viable quality control schemes have many sub-components, some monitoring the different aspects of the process - accuracy, precision, drift, linearity, carryover, while others compare the laboratory's performance between days. Obviously there must be a

balance between quality control and the clinical work requested of the laboratory. We shall propose later that the specific quality control scheme employed by a laboratory should be evolved from certain basic principles and be tailored to fit the particular laboratory.

Having examined the points raised by the question posed - 'how frequently should Quality Control be carried out?', let us now turn our attention to answering the question. There appears to be no reasonable nor practical way of carrying out continuous quality control on the measuring process. Procedures which we shall discuss in section 5.6 have been proposed. These procedures attempt continuous control by employing the actual patient's results. We shall show, however, that these procedures cannot afford the information required to control the measuring process. Thus if continuous monitoring cannot be achieved, quality control must be carried out at isolated instants of time. Just how isolated is the subject of future discussion. This will allow information of two types. Firstly, it gives information about the status of the measuring device at the particular instant of time at which the information was collected. Secondly, it should be possible by collecting such information to establish a 'picture' of the status of the process between control instants. We admit that between such control instants, we must assume that the status of the process can be represented by the information we have. Further, we note that specific specimens could be affected by 'one-off' lapses of control e.g. lack of proper mixing, too much or too little reagent etc. The instants between control information can be of any length and further, the control instants need not monitor all facets of the process.

Two extremes could occur. The process could be monitored at prescribed instants e.g. at the initial calibration in the morning, at mid-day and mid-afternoon, or the control instants could be spread uniformly throughout the day. We advocate the second approach, that is after the initial cleaning, washing, bleaching, calibration etc. of the machine/method, a uniform degree of control is maintained throughout the working day. This would appear to be the only reasonable action. If we do not accept this and opt for the first approach, a problem will arise if at the control instant an 'out of control' status is signalled. The only information available is the last check data which showed the process to be in control. Now, if only three quality control checks are made per day, this means that approximately one third or less of the laboratory's results could be in error. Further, there is no detailed information regarding the possible error, and hence we cannot reliably correct the suspected results. If this problem arises, when a uniform degree of quality control is maintained, we should have fewer results affected and also have information allowing for reliable correction to be carried out. As we have stated, the intervals between quality control checks can be of variable size. Normally, the intervals should be of a specified size, allowing for the clinical reliability required, but should be variable if required. Variability of the interval allows for a more rigorous control if required - perhaps because of a batch of special patient specimens, or perhaps because of special problems, electrical, hydraulic, manpower etc., which may occur infrequently during the working day.

The discussion of the actual frequency for the various subcomponents of quality control will be delayed for the present and

considered again in section 5.8. We shall see that it is important to consider the frequency of the different control specimens, 'blind' controls 'quality' control blood specimens and replicated patient specimens.

5.4 Control Charts

5.4.1 Introduction

Before we embark on a discussion of quality control in a haematology laboratory, we shall give a brief outline of two methods, which are currently used for presenting and analysing quality control information. The basic methods are quite general and we shall find that in discussing specific modes of control in later sections we will refer to these methods.

5.4.2 Shewhart Charts. (Standard deviation Plots)

In 1924 Shewhart suggested a method for investigating the quality of an industrial output, see Shewhart (B 32). Although the area of interest was industrial production processes, Shewhart's ideas have been extended and have become an important and frequently used procedure in quality control of many different types. In the following, we shall translate the fundamental ideas into a haematological context.

Shewhart maintained that variation in quality of output could be divided into two categories. That is the variation in patient results may be due to two sources. The first, is natural and random variation, i.e. the biological variation and the acceptable variability (imprecision) of the measuring process. The second, is what Shewhart called the variation due to assignable causes. In the haematological

situation these causes could be changes in accuracy, precision, linearity, carryover, drift, reagents etc. If the variation in the results reported by the measuring process can be attributed to natural and/or acceptable random variation, then the process is considered to be 'in control'. If, however, we cannot attribute the variation present solely to natural and/or acceptable random variation, then we must accept that some 'assignable cause(s)' contributes to the variation. In this case the process is deemed to be 'out of control'. Although we have accepted that there is some factor which can be 'out of control' which is affecting quality this does not necessarily mean that we can actually identify the physical cause of the loss of quality.

For the present, let us assume that the quality control information which we analyse comes from some quality control preparation(s). We shall discuss later what materials can be used.

Consider groups of measurements, usually sequential, made on the quality control material. Let the size of each of these groups be n. Two situations can arise. The group size, n, is small, i.e. less than sixteen, or the group size, n, is large, i.e. sixteen or greater. Consider a measuring process for a clinical parameter which has produced the following m groups of 'quality control values'. Between each of the groups, there would be sequences of patient specimens. The data are:-

$$(x_{11}, x_{12}, \ldots, x_{1n}), (x_{21}, x_{22}, \ldots, x_{2n}), \ldots, (x_{m1}, x_{m2}, \ldots x_{mn}).$$

In the simplest case, successive batches and successive measurements are independent. In practice, however, the successive errors in measurements may be correlated. The situation is made more

complicated when dealing with continuous measuring devices such as the Coulter-S. In such systems we require to consider the effects of carryover. If the measurements in each group are on the same quality control specimen, the effects of carryover can be overcome by running a 'blank' quality control specimen before each of the groups. Thus effectively each group is preceeded by a specimen denoted, x_{j0}, $j = 1, 2, \ldots, m.$, which is not included in any subsequent analysis. Now we shall consider how these data are handled when n is small and when n is large.

The group size, n, small.

Two quantities are calculated for each group - the group mean and the group range. The mean of the i^{th} group is denoted $\bar{x}_i$ where:-

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij} \qquad i = 1, 2, 3, \ldots, m.$$

This index is used to monitor the accuracy of the measuring process. The range of the i^{th} group, denoted r_i, is the difference of the largest and the smallest observation in the i^{th} group. The range reflects the precision (imprecision) of the process. In section 2.3.1, we indicated some effects of rounding in data. This is yet another situation in which rounding can affect the analysis. If, for example, we record, say 2-6 replicate measurements for MCV on a Coulter-S, commonly we find that exactly the same numerical value will be reported for each replicate. This may well lead to problems when estimating the standard error of the overall mean, see below.

If n = 1, i.e. only one control measurement is made at any specific instant, there is no information available concerning the precision of the process at these check times.

Denote by $\bar{\bar{x}}$, the overall group mean and let $\bar{r}$ denote the mean of the group ranges, i.e.

$$\bar{\bar{x}} = \frac{1}{m} \sum_{i=1}^{m} \bar{x}_i \qquad \text{and} \qquad \bar{r} = \frac{1}{m} \sum_{i=1}^{m} r_i$$

Commonly $\bar{\bar{x}}$ is taken to be the mean of the measuring process. Actually it is an estimate of this mean. Further, it may not necessarily be a 'good' estimate. This will only be true when our assumption concerning the stability of the measuring process while determining x_{ij} $i = 1,2,..,m.$, $j = 1, 2, \ldots, n.$, is justified. Frequently, some target value is used for $\bar{\bar{x}}$, e.g. the mean of the commercial preparation used to calibrate the process.

The average value of the group ranges, $\bar{r}$, gives a measure of the within group variability, i.e. the precision of the process. The quantity $\bar{r}$ can be used to estimate the standard error of $\bar{\bar{x}}$, $SE(\bar{\bar{x}})$ which is

$$SE(\bar{\bar{x}}) = \frac{\bar{r}}{d(n)}$$

Values for d(n) are given below in table 5.4

n	d(n)	n	d(n)
2	1.128	9	2.970
3	1.693	10	3.078
4	2.059	11	3.178
5	2.326	12	3.258
6	2.534	13	3.336
7	2.704	14	3.407
8	2.847	15	3.472

TABLE 5.4

A rough estimate for d(n) can be used by replacing the values in the table 5.2, by $\sqrt{n}$. The reliability of this estimate of the standard error of $\bar{x}$ decreases as n increases.

Group size large.

Two quantities are calculated - the group mean and the group standard deviation. The group mean x_i is calculated as in the above. The group standard deviation, s_i, is a 'better' estimate of the precision when n is large. To determine the standard error of $\bar{x}$, $SE(\bar{x})$ we note that:-

$$s_i^2 = \frac{1}{n-1} \sum_j^n (x_{ij} - \bar{x}_i)^2 \quad i = 1, 2, \ldots, m$$

$$s^2 = \frac{1}{m} \Sigma s_i^2$$

$$SE(\bar{x}) = \sqrt{\frac{s^2}{n}}$$

Irrespective of whether n is large or n is small, and assuming that $n \neq 1$, we determine the quantities $\bar{x}$, $SE(\bar{x})$, and $\bar{x}_i$ and either s_i or r_i. These quantities are sufficient to construct the Shewhart control chart for the mean of the measuring process.

Control Chart for the Mean.

In the original description by Shewhart, the following chart was suggested. We construct the so-called 'three-sigma interval' about the target value, which in this case is assumed to be the overall mean $\bar{x}$. This interval is given by $\bar{x} \pm 3SE(\bar{x})$. Having constructed this interval, a graph is constructed as shown below in figure 5.6, in which the group means, as they occur, are plotted.

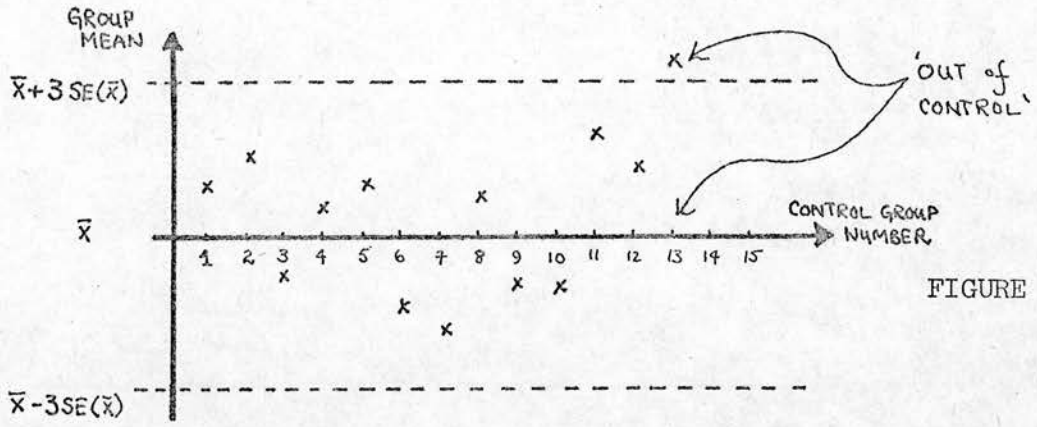


FIGURE 5.6

If a group mean falls outwith the three sigma lines, the process is considered to be out of control with respect to the mean. If the current group mean is within the three-sigma limits the process is considered to be in control.

It was shown by Dudding and Jenett (1942), that a modification to this procedure leads to better monitoring of the measuring process. They suggested that as well as constructing the three-sigma limits, the two-sigma limits should also be constructed. This interval is given by $\bar{\bar{x}} \pm 2SE(\bar{x})$. The three sigma lines are called the action lines while the two sigma lines are called the warning lines. If a group mean falls outwith the action lines, the process is stopped. If two successive means fall outwith the warning lines but within the action lines the process is also stopped. If one mean falls between action and warning lines, an immediate quality control group is measured. Otherwise the process is continued.

It has been suggested that a different type of inference can be made if two successive means fall outwith the warning lines but on different sides of $\bar{\bar{x}}$, namely that the process precision has altered.

Many other operating decision rules can be used. For example, if three consecutive group means fall outwith the interval $\bar{\bar{x}} \pm SE(x)$

stop the process. However, Wetherill (B 35) has made two points relating to these modified rules. Firstly, these modifications do not lead to a great improvement in the control of the process as compared to the action/warning scheme described by Dudding and Jennett (1942). Secondly, with such modification, the simplicity, a major attraction of the chart, is lost. Before we discuss the frequency of measuring control specimens, we discuss the construction of a control chart for the precision of the process.

Control Charts for the Precision.

Control charts can be constructed to monitor the precision of the measuring process. Often this is effected by plotting the group ranges. We note that in this case the minimum value for any of the $r_i, i = 1, 2, \ldots, m$ is zero. The control chart is constructed in a similar fashion with action and warning lines as before.

To construct a control chart for the range, i.e. to control the process precision, the following procedure is used. The target value is set as $\bar{r}$. To construct the action and warning limits for the control chart, two approaches are possible. Firstly, we could determine the standard error of the sample range and set the action/warning limits at $\pm$ 3/$\pm$ 2 standard errors from the target value. However this is not advised, even as a simplification to the following procedure, as it often results in negative lower limits. Secondly, we could use the percentage points of the studentised range (See Pearson and Hartly (B 31) Biometrika Tables for Statisticians). We have defined the studentised range in section 2.3.3 to be 'Range of sample/ σ'. To do this we determine the appropriate percentage points for the sample size used. For example if the sample size were three, the 1% point is 0.19, the $2\frac{1}{2}$% point is 0.30, the $97\frac{1}{2}$% point is 3.68 and the 99% is 4.12.

The action limits are then ($\bar{r}$ - 0.19 σ, $\bar{r}$ + 4.12 σ) and the warning limits are at ($\bar{r}$ - 0.30 σ, r + 3.68 σ). Commonly, only the upper action and warning lines are used, as shown in figure 5.7.

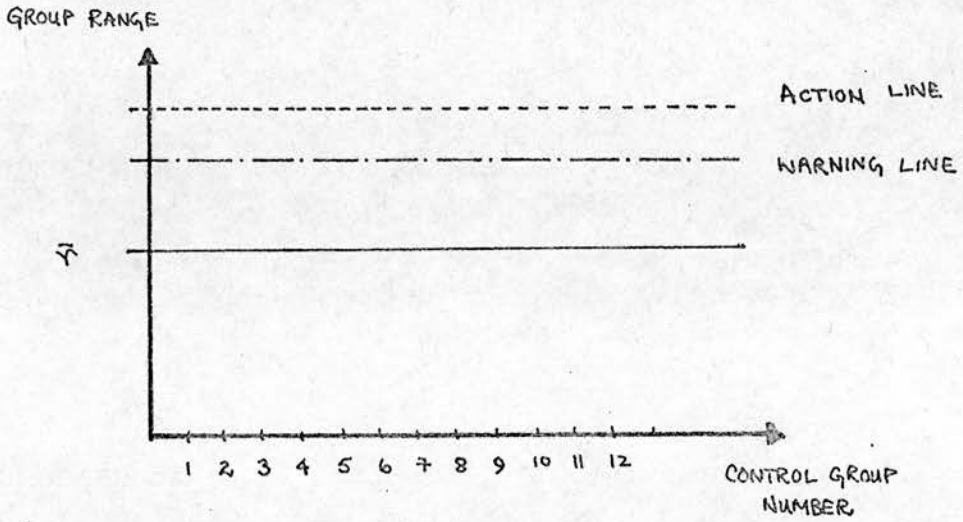


FIGURE 5.7

I think that it is important that the lower action and warning lines are also used in these circumstances. These lines reflect, possibly, if a process is 'too accurate'. This may in fact happen in laboratory situations, where technicians know when quality control specimens are being measured and further know what values are expected. We must remember that we wish to make reasonably objective statements concerning the reliability of the measurements of the clinical parameter.

Control charts for the precision of the process can also be based on the estimated population standard deviations obtained from the groups. In this case the percentage points, needed for setting the action/warning limits are obtained from tables of the F-distribution, by taking square roots. However, in practice and particularly in haematology we will rarely if ever have groups of sufficient size to justify the use of such charts rather than the range charts described above.

The main characteristic of the Shewhart chart is the ability to detect large or moderately large changes in the quantity being monitored quickly. However, these charts are not suited to detecting small changes quickly.

If we assume that the precision of the process, when in control, is σ, i.e. the standard deviation of replicate measurements is σ, and assume that the error in replicate measurements has a Gaussian distribution, we can determine, for any bias, the average number of control means before an out of control signal is given. This quantity is known as the average run length (A.R.L.). Obviously,we would wish for a scheme whereby for no bias, i.e. process in control, no out of control signal is given, and for any bias, the number of means plotted is zero or very small. However, this cannot be achieved without very great expense. For a bias of 0.4 σ, on average 200 means will be plotted before an out of control signal is given. If a bias of 0.8 σ, is present, the signal is on average after 71 means. For large biases, the signal will on average come much quicker, e.g. if a bias of 2.0 σ is present, on average six means will be plotted before the out of control signal is given. When no bias is present, that is the process is in control, an out of control will be signalled on average after 385 means have been plotted.

If both action and warning lines are used, the monitoring is better. This is due to the combined information from two sequential means.

A number of practical problems remain. Much of the literature and associated results require us to assume that the group means have Gaussian distributions. When n is large, this assumption can be justified as a result of the Central Limit Theorem, see section 2.2.2.

However, in practice when controlling a haematological measuring process, n will be small e.g. n = 1, 2, or 3. In these cases, the assumption that the group means have Gaussian distributions is reasonable since the individual replicate measurements can be assumed to have Gaussian distributions. Before we consider the frequency of the control measurements we note Wetherill (B 35).

'In many applications of quality control charts, samples of size 4 or 5 are used. The reason for this is partly that it is usually better to take a smaller sample size frequently than a larger sample size less frequently. A further point is that a small sample size keeps down the amount of arithmetic necessary for plotting.'

Both Duncan (1956) and Chiu (1973) have discussed the 'economic design of control charts'. The results which we give below follow their approach and relate to the simple control charts for the process mean with action limits only.

Let the group size be n, and let the group means have a Gaussian distribution with mean 0 and variance $\frac{\sigma^2}{n}$ when the process is in control, and a Gaussian distribution with mean either $-\delta\ \sigma$ or $+\delta\ \sigma$ and variance $\frac{\sigma^2}{n}$ when out of control. The action limits are set at $\pm\ k\ \sigma/\sqrt{n}$. The control measurements are made after h patient specimens. The problem is to choose h, k and n to give the control required.

The process is assumed to start in control and the number of specimens analysed until an assignable cause occurs is assumed to be a Poisson random variable with mean λ. Further the time taken to plot the control mean lets g patient specimens through the process.

If an assignable cause occurs between checks, the average number of specimens analysed is E(T) where

$$E(T) = \frac{\sum_{0}^{n} s \frac{e^{-\lambda}\lambda^{s}}{s!}}{\sum_{0}^{n} \frac{e^{-\lambda}\lambda^{s}}{s!}} = \frac{\lambda \sum_{0}^{n-1} \frac{e^{-\lambda}\lambda^{s}}{s!}}{\sum_{0}^{n} \frac{e^{-\lambda}\lambda^{s}}{s!}}$$

The probability of detecting an assignable cause is P, where

$$P = P(k,n,\delta) = \Phi(-k-\delta\sqrt{n}) + \Phi(-k+\delta\sqrt{n})$$

The probability that an assignable cause is detected r specimens after it has occurred is $P(1-P)^{r-1}$. Thus the average number of specimens measured is $\frac{1}{P}$. Thus the average number of specimens which are measured after an assignable cause has occurred and before an out of control signal is given is C, where

C = (Average time out of control until detection) + (Plotting delay)

$$= \frac{h}{P} - E(T) + g$$

The proportion of the time the process is in control is $\frac{\lambda}{C+\lambda}$.

Let us give some examples to show how this can be used to investigate certain control schemes, see table 5.6.5

TABLE 5.6.5

k	Average No. Specimens before cause λ	Frequency of Control Groups h	Group Size n	Bias δ	Plotting delay g	C	Proportion out of control %
2	40	20	2	1	2	53.96	42.60
2	40	20	5	1	2	15.62	71.91
2	40	40	5	1	2	33.89	54.14

There are many different computed quantities which could be used to compare different quality control schemes. We could investigate the Average Run length (A.R.L.) for different biases in the measuring process. From the above analysis, we see that the average number of specimens measured before a 'correct' out of control signal is given, C may also be used. The proportion of the time (i.e. the proportion of specimens) that the process is in control could also be used. We note that these computed quantities depend on the assumptions concerning the measuring process which we are willing to make and also on the parameter values which require to be estimated, e.g. the mean time to an assignable cause. Thus, the analysis gives only rough estimates of the computed quantities. The situation is made very much more complicated if we were to assume that action and warning limits were used, or if we were to assume that the bias present in the measuring process was variable.

The numerical values which we have presented, give indications as to the quality control which should be imposed on the process. In practice, in the haematological situations which we are investigating, the group number, n, will be small e.g. n = 1, 2, or 3. From the table presented we see that the process should be monitored frequently. However, the numerical values indicate a situation worse than we actually expect in practice. Commonly the action/warning lines are used and this improves the control and in general lessens the frequency of monitoring. In section 5.8, we discuss a scheme which is currently used in a haematology laboratory. This discussion will include comments on the affects of frequency of monitoring.

For the present we turn our attention to an alternative method of presenting the quality control data which has been collected.

5.4.3 Cusum Control Charts.

We have noted that Shewhart charts have the disability of not being able to signal small changes quickly. The use of a Cusum (Cumulative Sum) technique, however, can overcome the problem of signalling small biases quickly. The Cusum technique uses all past information from the initial set-up of the measuring process or from the last recalibration.

We assume that exactly the same information is available to us as when discussing the Shewhart control charts, i.e. m groups of n replicate quality control measurements. Once again, we calculate the group means $\bar{x}_i$, $i = 1, 2, \ldots, m$ and the overall mean $\bar{\bar{x}}$. Further, depending as before on the size of the groups, n, the group ranges, r_i, or the group standard deviations, s_i, are also calculated. For the present we shall discuss the construction of a cusum chart for the process mean.

Cusum Chart for Process Mean

As before a target value is set for the quality control specimens. This will either be the overall mean $\bar{\bar{x}}$, of some set of quality control data, or a prescribed value, e.g. the assay value of a commercial product. Let us denote the target value by k. The basis of the cumulative sum technique is the cumulative or partial sums S_r, where

$$S_r = \sum_{i=1}^{r} (\bar{x}_i - k) \qquad r = 1, 2, 3, \ldots.$$

These sums are the cumulative sums of the deviations of the group means from the target value k. We note that when the process is in control, i.e. $E(\bar{x}_i) = k \quad i = 1, 2, \ldots.$

$$E(S_r) = 0$$

Thus the graph of S_r against r should show only small random scatter

about zero. However when the process is out of control, we have

$$E(S_r) = E(\Sigma \bar{x}_i) - rk$$

If there is a constant bias in each group mean, such that $E(\bar{x}_i) = \mu$ then

$$E(S_r) = r(\mu - k)$$

Thus if we plot the partial sums S_r against r, the outcome should give a line with gradient $\mu - k$. Because we are interested in the slope of the graph, it is important in practice to take care in choosing the scale of the graph. Ewan (1963) has discussed this problem. It is common to choose a scale which has one unit in the hotizontal axis, i.e. group number, equivalent to 2 σ units in the vertical axis, i.e. group means, where σ is the precision of the process when in control.

Some early papers describing cusum techniques gave significance tests for slope change, i.e. possible bias in the measuring process, - see Page (1954) and Ewan and Kemp (1960). These methods could be programmed for computer usage. However, Barnard (1959) gave a graphical method for interpreting the cusum graph, which is equivalent to these numerical procedures.

Barnard suggested that a 'V-mask' be placed on the cusum chart, so that the current group mean coincides with a special point on the V-mask. If the graph cuts either limbs of the mask, the process is deemed to be out of control. Figure 5.8 shows a cusum graph of a process which is in control at time t_o. Figure 5.9 shows a process which has experienced an upward shift in the process mean. Figure 5.10 shows a process which has experienced a downward shift of the process mean.

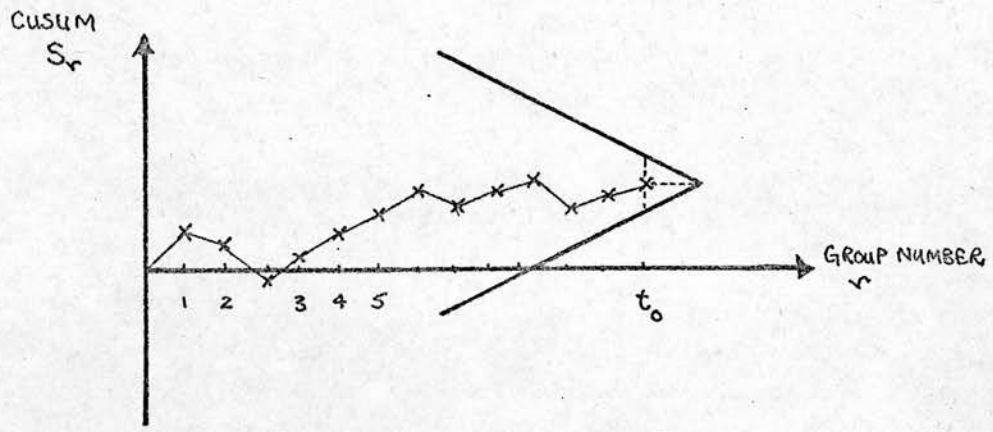


FIGURE 5.8

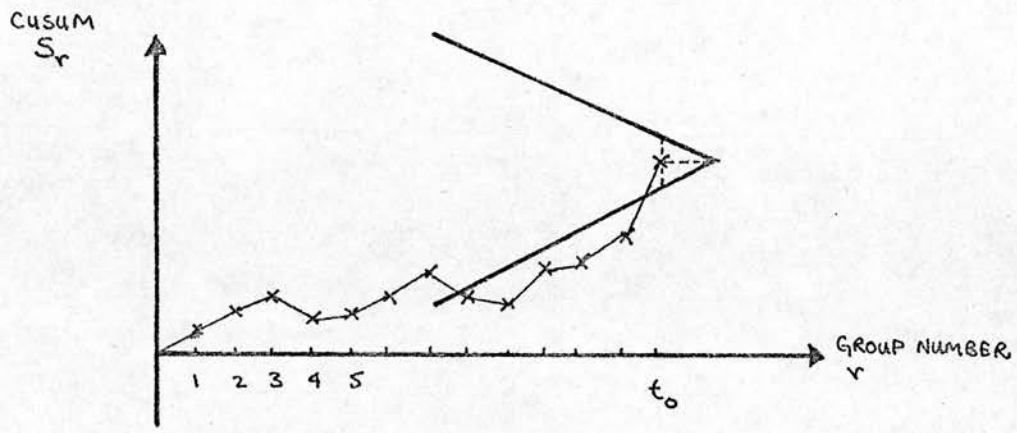


FIGURE 5.9

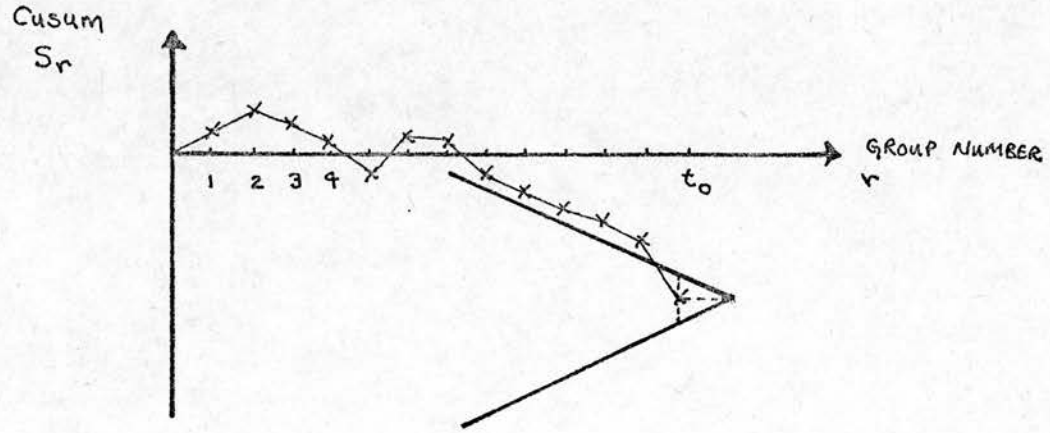
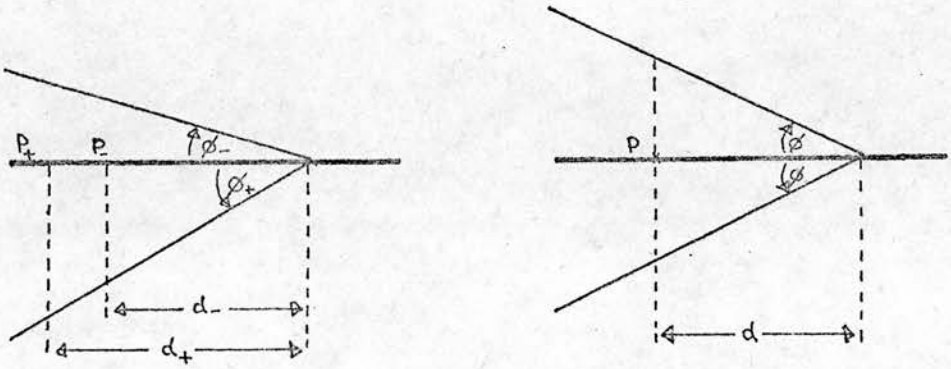


FIGURE 5.10

The construction of the V-mask (symmetric) depends on two parameters d and ϕ. In cases where we wish to construct asymmetric V-masks, perhaps to monitor a process which is more likely to have a bias in a particular direction, four parameters may be necessary,

d_+, d_-, ϕ_+ and ϕ_-. Examples of both types of mask are given in figure 5.11. The points P, P_+ and P_- are the points on the V-mask which should be coincident with the last group mean plotted.

FIGURE 5.11



These parameters, either d and ϕ or d_+, d_-, ϕ_+ and ϕ_- specify the cusum tests completely. In the examples of V-masks which we have given, straight edges have been used. This need not be the case. Other masks can be used e.g. parabolic masks. Such masks have slightly different properties from the straight edged masks. However, most authors consider any improvement in the monitoring of the process to be negligible.

Johnston (1961) showed that this cusum technique can be considered as a 'backward' sequential process. He has shown the relationship between the parameters of the cusum mask and the parameters of a sequential technique as explained by e.g. Armitage (B 3). For the practical details concerning the construction of the V-mask see Woodward and Goldsmith (B 37).

A number of practical problems arise in using this method. The use of the V-mask in the laboratory becomes less strange and disconcerting after a short introductory period. However, the graphing procedure which we have explained can take up a substantial space, which is

usually at a premium in the laboratory. This practical problem has been solved by using a 'Decision interval scheme', - see Ewan and Kemp (1960), Ewan (1963) and De Bruyn (B 11). It can be shown that the V-mask schemes and the decision interval schemes, which do not necessarily need graphing procedures, are equivalent.

A further problem, which was also experienced when discussing Shewhart charts, concerns the assumption that the group means have a Gaussian distribution. We can, as before, accept that this is a reasonable assumption to make.

A similar type of analysis to Duncan's has been presented by Taylor (1968) concerning the economic design for a Cusum control scheme. The analysis is more complicated and will not be discussed here.

Let us compare these different methods, Shewhart charts and Cusum charts, of presenting quality control information. Both methods, as we have noted, can be made visually appealing; not a small consideration in a busy laboratory. In practice, the Shewhart chart is simpler to use, and for the same number of false alarms as the Cusum, it can be made to react more promptly to large biases.

For the same number of false alarms as the Shewhart chart, the Cusum can be made to react more promptly to small biases. When an out of control signal is given, the time at which the process went out of control can be estimated from the graph. Further, an estimate of the bias can easily be obtained from the graph. Since both large and small biases and experienced in a haematology laboratory, it may well be that both schemes should be used simultaneously. (Any attempts to combine the two methods into one chart has resulted in complicated impractical schemes.) However, the calculations and graphing of two

control charts will commonly be prohibitive with respect to time.

Finally, we remark that in using both Shewhart and Cusum charts, the underlying distribution, particularly if small non-replicate control groups are used, can affect the properties of the charts. It is assumed throughout, that the points plotted have Gaussian distributions. Some attempt has been made to construct Distribution-free (non-parametric) charts e.g. see McGilchrist and Woodyer (1975).

5.5 Estimation of the Process Mean

5.5.1 Introduction

In using a quality control scheme, much interest is centred on monitoring the process mean. In section 5.4, we have described methods for monitoring the process accuracy by analysing groups of quality control specimens. Many authors are critical of the use of special quality control specimens. Two common reasons given are:-

(i). Quality control specimens are not subject to the environmental factors which may affect the patient samples. Thus, the quality control is based only on the quality of the measuring process in close proximity to the particular measuring device.

(ii). Experienced technicians, despite efforts to make the quality control scheme 'blind', can spot control specimens and may take more care with them.

A number of authors have suggested methods for using the complete record of measurements made by the measuring process, to monitor the reliability of the process. These methods, instead of monitoring the accuracy by collecting information on quality control specimens,

are designed to monitor the accuracy of the process by investigating the mean of the patient population from which the specimens are drawn. We shall discuss in section 5.6 formal quality control schemes which use such methods. For the present we investigate indices which have been suggested for monitoring the patient population mean.

While dealing with quality control schemes based on special quality control specimens, we made the assumption that the measurement error had a Gaussian distribution. Although this assumption will not always be justified, it has been found to be reasonable. A major problem is encountered concerning the distributional properties of the patient population - not just the particular underlying distribution but also the sampling characteristics, which will not usually be random.

A further problem, is similar to the problem of outliers which we have discussed in sections 2.3.3-5. In that discussion of outliers, we dealt with the problem of 'wild/rogue' observations in replicate measurements. The replicate measurements on a single specimen should come from the same parent population. (We assumed a Gaussian distribution with mean equal to the 'true' specimen value and standard deviation equation to the precision of the measuring process while in control.) Outliers, if present, are assumed to be observations from some other distribution with different mean and/or variance. The outlier problem which now faces us, is concerned with what may be called 'biological outliers'. When considering a clinical parameter, the average value is usually reported. Often it is the midpoint of the 'normal range'. In fact this reported value should carry with it some description of the population it attempts to

describe, which is assumed to be the population of 'normals'. In practice, even the population of 'normals' will comprise different sub-populations, which relate to sex/age etc. differences. Thus, for different pathological states the clinical parameter might have values far removed from the average clinically accepted value, see section 5.2. It has been proposed that to estimate the accuracy of the measuring process this clinical average value is monitored. However, the mean of a group of patient values, especially if small, can be seriously affected by these 'biological outliers'. This effect can be marked if, as happens in routine situations, the specimens presented for measurement arrive in batches, or isolated pathological specimens become mixed up with clinically 'normal' specimens. As before, if we wish to estimate the 'true' population mean with relatively few specimens in any batch, we should try to eliminate the effect of these biological outliers. (These outliers may be caused by a freak result of the measuring process and hence not a 'biological outlier' in the sense above. However, there is no method other than comparing concomitant clinical results to establish if this is a 'biological outlier'.)

5.5.2 Proposed Estimators.

For the present we discuss indices which could be used to estimate the mean of the patient population - i.e. the average of the clinically 'normal' patient population. Consider a measuring process for a specific clinical parameter, which produces batches of results:-

$$(x_{11}, x_{12}, \ldots, x_{1n}),(x_{21}, x_{22}, \ldots, x_{2n}),\ldots,(x_{m1}, x_{m2},\ldots,x_{mn}),\ldots$$

These measurements are actual patient measurements. We denote by μ the mean of the 'normal' patient population. Bull, Elashoff, Heilbron

and Couperus (1974) have discussed the use of the following estimators of the population mean.

(i). The Sample Mean

Denote by $\bar{x}_i$ $i = 1, 2, \ldots,$ the mean of the i^{th} batch of patient results.

$$\bar{x}_i = \frac{1}{n} \sum^{n} x_{ij} \qquad i = 1, 2, 3, \ldots.$$

As we have stated in section 2.2.2, providing that n is large, then for almost all reasonable distributions for x_{ij}, the distribution of $\bar{x}_i$ is approximately Gaussian. In some situations, however, the distribution may be skewed and/or highly dispersed. If the distribution of x_{ij}, including pathological values, is approximately symmetric, then the overall $\bar{x}$ and the batch means $\bar{x}_i$ will be reasonable estimates of the population mean.

$\bar{x} = \frac{1}{N} \sum^{N} \bar{x}_i$, where N is the total number of patient batches considered.

We monitor the accuracy (i.e. the population mean) using $\bar{x}_i$, i>N, setting the target values as $\bar{x}$.

This scheme has a number of inherent problems. As before we require to assume that the measuring process was in control while producing the target value $\bar{x}$. Further, we require to assume that the precision of the measuring process is constant not only for all batches, but also for all values of the clinical parameter. The main problem concerns the biological outliers discussed above. The sample mean is affected by the presence of biological outlier(s).

(ii). A Moving Average, $\bar{x}_{ma}$

By using the current batch mean, $\bar{x}_i$, to estimate the population

mean, we do not take into account recent information concerning the accuracy of the process.

Consider utilizing the information contained in the immediate preceding batch mean. We can achieve this by constructing a moving average $\bar{x}_{ma,i}$, for the current batch. This average has the form:-

$$\bar{x}_{ma,i} = r\,\bar{x}_i + (1 - r)\bar{x}_{ma,i-1} \qquad i = 1, 2, \ldots\ldots$$

$$0 < r \leq 1 \text{ and } x_{ma,0} = A$$

The value of A will commonly be the value of the population mean, as recorded for the measuring process in the past, when the process is in control. The choice of r depends on the specific situation. It will depend on the initial value A, the precision of the measuring process when in control and the bias which is required to be detected. A control chart has been suggested for this type of average - see Roberts (1959). (This average is optimal for controlling a first order auto-regressive process.)

This estimator suffers from the same problem as the mean $\bar{x}_i$. That is it is affected by biological outliers.

(iii). The sample Median x_{med}

In an attempt to minimise the effects of biological outliers, we could estimate the population average by the batch median $x_{med,i}$. This estimate will estimate the population mean only if the distribution of patient specimen values is approximately symmetric. Irrespective of the distribution, but under certain regularity conditions, the sample median has an asymptotically Gaussian distribution with mean equal to the population median and variance equal to $4nM^2$, where M is the population median.

An unattractive property of this estimator lies in the amount of potentially useful information which is not used. When the population is truely Gaussian, the efficiency of this estimator as compared with the batch mean is approximately 64%. Bull, Blashoff, Heilbron and Couperus (1974) have suggested that a moving batch median could be used. This estimate, $\bar{x}_{med,i}$ has the form:-

$$\bar{x}_{med,i} = r\, x_{med,i} + (1-r)\bar{x}_{med,\, i-1} \qquad i = 1, 2, \ldots, 0 < r \leq 1$$

where $x_{med,i}$ is the median of the i^{th} batch and $x_{med,0} = A$. The value of A is obtained from past records.

In practice, the batch median will not give as good an estimating procedure as some of the following estimators. Thus, it would seem to be a waste of effort to carry out the arithmetical computations etc., to produce the moving batch median.

(iv). The trimmed Mean. $\bar{x}_t$

This average attempts to reduce the effects of 'biological outliers' by eliminating some of the most extreme measurements in the batch. The measurements of the i^{th} batch are arranged in ascending (or descending) order of magnitude, giving:-

$$z_{i1} \leq z_{i2} \leq z_{i3} \leq \ldots\ldots \leq z_{i\, n-1} \leq z_{i\, n}.$$

We discard, say $s \leq n/2$ of these values from either end of this ordered sequence. The trimmed mean of the i^{th} batch is defined to be:-

$$\bar{x}_{t,i} = \frac{1}{n-2s} \sum_{j=s+1}^{n-s} z_{ij} \qquad i = 1, 2, 3, \ldots.$$

David (B 10) has discussed the robustness of this estimator,

and notes that since $\bar{x}_i$ is asymptotically Gaussian, then $\bar{x}_{t,i}$ must also have this property.

A moving average trimmed mean can be constructed in an obvious way. Johnson and Kotz (B 22) give the relative efficiencies of the trimmed mean in comparison with the mean $\bar{x}_i$. They also remark that the Winsorised mean (see (v)) is more efficient than $\bar{x}_{t,i}$ given that the underlying distribution is Gaussian.

(v). The Winsorised Mean $\bar{x}_w$.

Using the same notation as in (iv) above, the Winsorised mean is constructed by replacing the values $z_{i1}, z_{i2}, \ldots, z_{is}$ by $z_{i(s+1)}$ and replacing the values $z_{i(n-s+1)}, \ldots, z_{in}$ by $z_{i(n-s)}$. The Winsorised mean for the i^{th} batch is defined to be:-

$$\bar{x}_{w,i} = \frac{1}{n}\left(s\, z_{i(s+1)} + s\, z_{i(n-s)} + \sum_{j=s+1}^{n-s} z_{ij}\right)$$

This average is an attempt to produce an estimate which is not affected by extreme batch values, which possibly are biological outliers. This is done by replacing the most extreme values in the batch by less extreme values. Thus, some acknowledgement is made of the fact that the discarded values may contain valuable information relating to the population mean. (See comment in (iv).)

(vi). A Moving Average Due to John $\bar{x}_J$

This average is described fully by Bull et al. (1974). To construct the average, the batch measurements are ordered as in (iv) and (v). The extreme measurements z_{i1} and z_{in} are discarded. Of the remaining ordered measurements, the 'middle' k are summed to give $S_{m,i}$. The remaining n-k-2 measurements are

summed to give $S_{o,i}$. A weighted average is constructed from these sums giving $\bar{x}_j$ as

$$\bar{x}_{J,i} = \frac{W_m S_{m,i} + W_o S_{o,i}}{(n-2)(W_m + W_o)} \quad \text{where } W_o + W_m = 1.$$

This average simply infers that the more extreme measurements are considered to be less reliable and should contribute less to the average.

(vii). An Alternative Moving Average $\bar{x}_B$

Bull et al have suggested the following moving average, which has the form,

$$\bar{x}_{B,i} = (2-r)\, \bar{x}_{B,\, i-1} + r\, d_i \qquad i = 1, 2, 3, \ldots.$$

They consider in particular, $r = 1$ and indicate that the form of d_i which gives 'best' results is:-

$$d_i = \text{sgn} \left(\sum_{j=1}^{n} D_{ij} \right) \left(\sum_{j=1}^{n} \frac{D_{ij}}{n} \right)^2$$

where $D_{ij} = \text{sgn}(x_{ij} - \bar{x}_{B,i-1}) \, |x_{ij} - \bar{x}_{B,i-1}|^{\frac{1}{2}}$

As we shall comment in the following discussion, this average entails a complexity of arithmetical procedures etc., which will be an unacceptable constraint in the routine laboratory.

(viii). The Sample Mode x_m

For completeness, we include the sample mode x_m. We shall discuss the use of this measure of location in a particular method of control, see 5.6.4. The sample mode, basically, is the sample value which occurs most frequently. However there seems to be some dubiety in the actual definition of this quantity. As Armitage (B 2) indicates, the mode relates to a histogram or frequency distribution of a set of observations,

and in fact 'any value of the variable at which the frequency function curve reaches a peak is called a mode'. Thus, although the median and mean are always unique, this is not necessarily true of the mode. Yule and Kendall (B 38) point out that -

'It is, in fact, difficult to determine the mode for such distributions as arise in practice, particularly by elementary methods. It is no use giving merely the mid-value of the class-interval into which the greatest frequency falls, for this is entirely dependent on the choice of the scale of class-intervals. It is no use making the class-intervals very small to avoid error on that account, for the class-frequencies will then become small and the distribution irregular. What we want to arrive at is the mid-value of the interval for which the frequency would be maximum, if the intervals could be made indefinitely small, and at the same time the number of observations be so increased that the class-frequences should run smoothly.'

Even Pearson in 'The Grammar of Science' (B 30) acknowledges the practical difficulties associated with the practical use of the mode.

5.5.3 Comments on The Estimators of the Population Mean

It will be obvious, that the more information is collected and used concerning the accuracy of the measuring process, the better should be the monitoring of the process. Further, some methods of handling data are more efficient, depending on the situation. Thus we expect some of the estimators (i)-(viii) to perform better than others. No single one of the estimators is best under every possible

set of circumstances,

Let us consider those situations in which these estimators may be used. The personnel, who are likely to be in charge of supervising the calculation of such quantities, are not usually statisticians. These technicians have often many other tasks to undertake while carrying out this quality control supervision. Further it is often the case that there is no laboratory facility, be it calculator, computer or computer link, to help with the tedious arithmetical computations required for most of the quantities presented above.

We note that most of these estimators involve some computation and/or time to organise the batch measurements. There is no suggestion that these operations are beyond even the most junior technician, but when all the other duties of the technician are taken together in the atmosphere of the laboratory, it would be surprising to find that such work was carried out error free. We remark that it would appear to be much better to implement a slightly less efficient estimator which requires little time than an estimator which is more efficient but is complicated to use.

Bull et all (1974), after some trials, suggest that the 'best' values for r and N were 0.4 and 20 respectively for the indices $\bar{x}_B$ and $\bar{x}_{ma}$, and a moving average constructed from $\bar{x}_J$ to perform best. The authors had given $\bar{x}_B$ a most extensive investigation in routine laboratories. A point made at the end of their paper, seems to be important. They suggest that if calibration control at the 2-3% level is desired then the population coefficient of variation should be less than 10%. (The population coefficient of variation is the ratio of the population standard deviation to the population mean.)

We shall see in the following section that quality control based on the results of measurements made on patient's specimens is fraught with difficulty. To justify the use of the estimators, we require the basic assumption that the patient's specimens can be considered to be a random sample from the population under consideration. That is, the order that the specimens are presented to the measuring process can be assumed to be as if a random sample from the population was being presented. As we show later this assumption is very rarely justified. For this reason we would discourage the use of such estimators in the monitoring of the process accuracy. However, procedures have been proposed which employ some of these estimators, and for this reason we shall discuss these procedures in more detail.

5.6 Formal Quality Control Schemes using Patient Data.

5.6.1 Introduction

In section 5.5.1, we indicated that many practitioners argue that special quality control preparations cannot control the measuring process completely. Frankel and Ahlvin (1967) have put it in the following way.

'The regular and frequent use of pooled or commercial sera is the most common method of quality control presently employed. Sole dependence on this method, however, has obvious deficiences. The samples are known to be controls by the laboratory personnel and may be handled with greater care than routine specimens and, because these samples are not whole blood drawn from patients, they are not completely

comparable to routine specimens. In addition, deterioration or improper preparation of the test sample may yield false values. Thus, additional means to ensure control of laboratory procedures are continuously being sought. It is the opinion of many laboratory directors that the most suitable method would be one of an internal type, whereby the routine laboratory results are employed as the measure of control.'

Taking this last point made, we propose to discuss three such methods, which have been suggested.

5.6.2 The Number Plus Method

Hoffmann and Waid (1963) proposed the following quality control scheme, which is suggested to be less time consuming than using reference materials. It is proposed that this scheme is used to augment existing quality control.

Consider a measuring process for a particular clinical parameter. A large number of measurements (it is suggested in excess of 500) have been made on the population of interest. This information is used to construct a histogram which has the possible form as shown in figure 5.12.

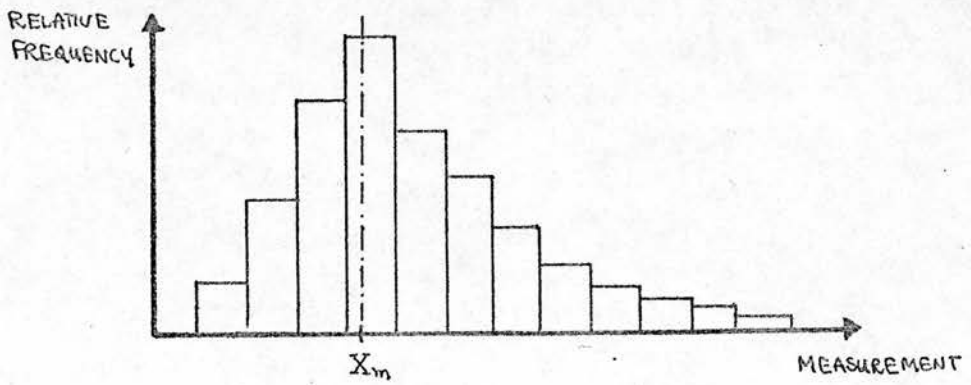
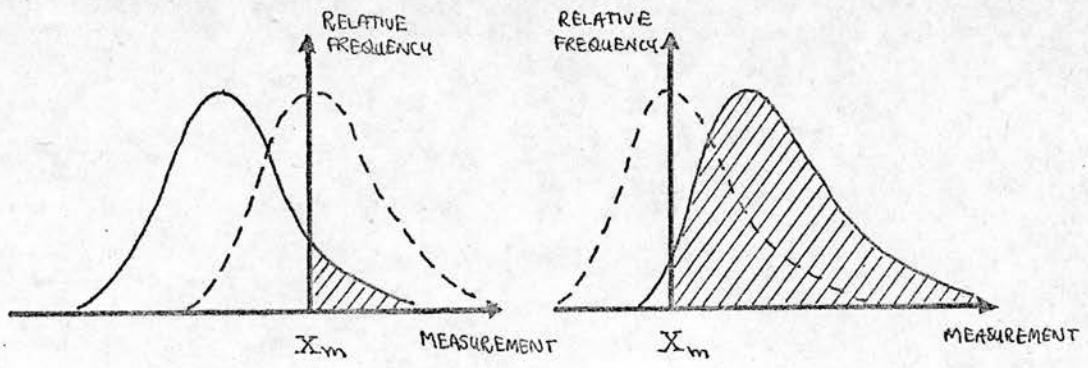


FIGURE 5.12

The histogram will hopefully have a single peak i.e. be unimodal. The implications are that clinical parameters which result in multimodal

or 'flattish' histograms are unsuited to this method of control. The clinical value corresponding to this maximum is the mode x_m, see section 5.5.1(viii). In practice, the mode will be situated within the 'normal range', however that be constructed. If the measuring process is stable and if the population in which we are interested is also stable, the distribution, and the corresponding histogram of patient's results will be stable with respect to shape and location. From the constructed histogram, the percentage of patient results greater than x_m can be estimated, say 100p%. Then p is an estimate of the probability that a specimen chosen at random will result in a value larger than x_m.

The patient results are inspected sequentially, and each specimen producing a value greater than x_m is given a 'plus'. When the measuring process is in control, the expected percentage of specimens receiving a 'plus' will be 100p%. (We note that this is not strictly true, since p is an estimate.) If the accuracy of the process changes, this expected percentage will be increased or decreased, see figures 5.13/14.



Downward drift

FIGURE 5.13

Upward drift

FIGURE 5.14

To operate the proposed quality control scheme, we monitor the proportion of specimens receiving a 'plus'. This is done by a modified Shewhart chart. The chart is for a Binomial random variable and not for a continuous random variable (Gaussian) as discussed in section 5.4.2. Briefly, a batch of n patient specimens is taken and the number of pluses, r, is determined. The random variable, r, is plotted on a Shewhart chart with target value np and action limits at the '95% confidence limits for the number of pluses when the process is in control', i.e. $np \pm 2.0 \sqrt{n\,p(1-p)}$. We note that these confidence limits are not exact, depending on the assumption that for large n the Binomial variate has an approximate Gaussian distribution. Hoffmann and Waid (1963) proposed that batches of size 50 should be used.

A number of points should be noted concerning the inherent assumptions which have been made. It is assumed that, if the measuring process is out of control, only a change in location will be experienced, see figures 5.13/14. It is possible however, that a shift in location and/or a change in distribution shape is experienced when the process is out of control. If a change in both location and shape happens simultaneously, it may be possible for the target value of np to be realised, but for the process to be out of control, see figure 5.15.

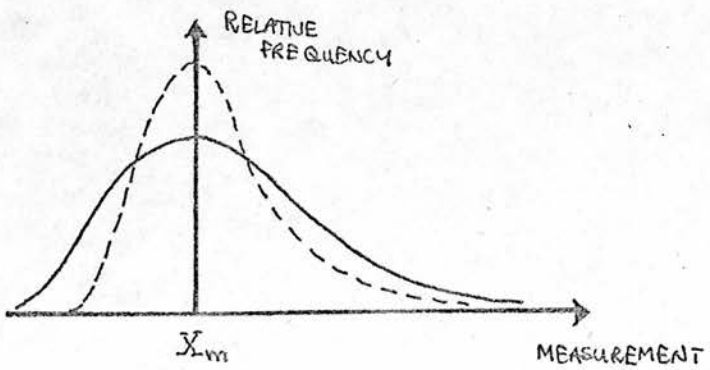


FIGURE 5.15

To construct the 95% confidence interval for the number of pluses when the process is in control, Hoffmann and Waid follow the accepted convention that since the batch size is large, 50, a Gaussian approximation can be employed. We note that this, on average, will result in inaccurate limits. Any attempt to construct exact limits is not advised due to other problems affecting this procedure.

For the statistical procedures employed to be justified, the specimens must be a random sample from the population of interest. This will rarely occur in practice. The specimens often arrive in batches from special clinics, wards etc., and this will probably produce periodic bias. Further, abnormal results, outliers and biological outliers, will often be presented for repeat analysis, thus producing in practice 'heavy' tails in the distribution of patient results. Frankel and Ahlvin (1967) in their evaluation of the method, used truncation limits in an attempt to eliminate the effect of these repeat determinations. They found that in the evaluation, covering two hospitals, the assumption regarding a 'stable population' was not justified. Their inference is that this instability causes too many 'false alarms'. These findings support the conclusions of Peenen and Lindberg (1965). However, these authors blamed the asymmetry of the patient population. This reason, obviously worthy of note, is probably less important than the non-randomness of the sampling. Amador, Hsi and Massod (1968) have carried out a further investigation, suggesting that the method 'does not provide the sensitivity to errors which is necessary for the proper conduct of a clinical laboratory'.

5.6.3 The Method of 'Average of Normals'

Hoffman and Waid (1965) have proposed another method of quality control based on the laboratory measurements of patient specimens. They have called this method the 'average of normals'. This method is frequently used in laboratories, (Ch.6, table 9b) . However, it is more than likely, that each laboratory has modified the procedure which we give to 'fit' their own laboratory.

Let us assume that the accuracy of a specific clinical measuring process has to be controlled. Once again a histogram of past patient results is constructed. A range (x_l, x_u) is defined to be the 'normal range' for the clinical parameter. The patient results reported by the measuring process are investigated sequentially and are discarded, for control purposes, if they do not lie within the interval (x_l, x_u). The results which are retained are averaged in groups of size n. Hoffman and Waid suggest that n = 10. These means are plotted on a Shewhart chart (see section 5.4.2).

By using this method, we obviously delete extreme patient results, outliers both biological and process mistakes. To some extent this overcomes a criticism of the last method. Statistically, we are operating a quality control scheme based on censored data. The process is deemed out of control if these censored means deviate substantially from the 'target value' whose construction has been discussed previously.

The use of a 'normal range' is not necessary. All that is required is some interval which contains a reasonable proportion of the underlying population, when the process is in control. Leclercq (1975) has discussed the selection of the optimal 'truncation limits'

in order to obtain simultaneously 'the lowest variability in the (daily) mean and the highest recovery of the introduced error'. If a reasonably large percentage is not used, the control scheme is not sensitive to shifts in the accuracy.

On the other hand, we have already noted the problem associated with too large a percentage.

Assume that the underlying distribution is Gaussian with mean μ and variance σ^2 and represent by $\phi(z)$ and $\Phi(z)$ the probability density function and the cumulative distribution function of the Standard Gaussian random variable. Let the 'normal range' be (a,b). The mean of the censored sample is denoted $\bar{x}_{c,n}$. Then we note that:-

$$E(\bar{x}_{c,n}) = \mu - \frac{\sigma\phi(\frac{b-\mu}{\sigma}) - \sigma\phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$$

$$V(\bar{X}_{c,n}) = \frac{2}{n}\{1 - \frac{(\frac{b-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma}) - (\frac{a-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} - \{\frac{\phi(\frac{b-\mu}{\sigma}) - \phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\}^2\}$$

If the 'normal range' is symmetric about the population mean, then we have:-

$$E(\bar{x}_{c,n}) = \mu$$

$$V(\bar{x}_{c,n}) = \frac{\sigma^2}{n}\left[1 - \frac{2(\frac{b-\mu}{\sigma})\,\phi(\frac{b-\mu}{\sigma})}{2\,\Phi(\frac{b-\mu}{\sigma}) - 1}\right]$$

Further the skewness of the sampling distribution is zero. We note that the standard error of $\bar{x}_{c,n}$ is always less than that of the uncensored mean of the same sample size. Thus action and warning limits for the Shewhart chart must be constructed from information relating to these censored data.

Once again, many of the problems identified with the last method afflict this one. Amador, Hsi and Massod (1968) have evaluated this method of quality control, reaching the same conclusion as that given for the previous method.

5.6.4 A Modal Control Procedure

A third method of quality control using patient results, has recently been proposed by Pragnell and Johnston (1976). The method uses the mode as the index for monitoring the accuracy of the process. The authors have used this quantity i.e. the mode, to combat the criticism that the mean value of the batches is affected by the large number of abnormal values which are measured (see Cavill (1971)).

The machine the authors specifically consider, the Coulter-S, is cleaned and washed through with Isoton. It is primed with blood at least ten times. A blood sample from a single donor is counted ten times and the results are used to establish whether the precision of the machine is acceptable. Acceptable ranges for the precision of each channel are given by the authors.

For each specimen counted in routine analysis, the white blood count WBC, the mean cell volume MCV, the mean cell haemoglobin MCH, and the mean cell haemoglobin concentration MCHC are recorded on a frequency distribution, truncated to the whole number below. It is suggested that after twelve results have been recorded, the modal value for MCH and for MCHC can be obtained. However, the WBC and the MCV may require a greater number of results. (The authors suggest using a 'best fit' line to estimate the mode, but without further explanation one must conclude that this has no meaning.)

It is suggested that the process is in control if the modal values fall within the ranges given below in table 5.6.

TABLE 5.6

	Acceptable Range		Minimum Change Detectable
WBC.	7.5 ± 1.0	10/1	0.5
MCV.	87 ± 1	fl.	1.0
MCH.	29.5 ± 0.5	pg.	0.25
MCHC.	33.5 ± 0.5	g/dl	0.25

The suggested minimum changes which can be detected are also given in table 5.4. If an out of control signal is given, a check list is used to correct any possible error(s). Among the points on this corrective action is a check of the accuracy of the haemoglobin channel.

By way of justifying this procedure, the authors show that over the past years while this method has been used, the laboratory's performance, as measured by the DHSS (LDAG) and BCSH Haematology Quality Control trial has usually been better than acceptable. (See chapter 6). Of course, as we point out later, this does not necessarily guarantee the quality of the laboratory's work from day to day. We shall see that many laboratories produce 'better than average results' which do not reflect the true status of the laboratory.

Firstly, we should note that the frequency distributions given as examples by Pragnell and Johnston, contain many more than the suggested twelve measurements e.g. for MCH the examples given have 22, 23, 23, 23, and 22 measurements respectively. Thus, despite their claims it would appear that we require a substantial amount of

information before we can identify the batch mode. We can show that the mean will signal out of control, on average before the mode. We show this for the example given by the authors using a Shewhart chart. The example concerns the MCH channel, and uses the truncated data presented in table 5.7.

TABLE 5.7

Batch	I	II	III	IV	V
Mode	29.5	29.25	29.00	28.75	28.50
Number in batch	22[6]	23	23	23	22
Mean	29.50	29.34	28.89	28.76	28.50

If we construct 'action lines' for the Shewhart chart using the standard deviation of the biological population to be 1.5, as reported by Coulter, then these will be (28.9, 30.1). We see that out of control would have been signalled by batch III, whereas it would be batch IV if the mode were used. The slowness of the mode to signal an out of control does not prevent a great number of false alarms. Leclercq (1975) has given an example comparing the batch mean, median and mode in control situations. The minimum batch size was 20. He comments, 'it may be quite difficult to obtain an exact definition of the mode. Furthermore, that in most instances the variability of the mode is greater than that of the mean'. He dismisses the use of the mode as do other participants of the symposium at which the paper was read.

The use of the sample mode in statistical analysis is rare. The distributional properties of the mode are complex and very little

discussion has been reported on the statistical properties of schemes such as that devised by Pragnell and Johnston. I do not think that this procedure is as good as other schemes which we shall discuss, and very probably it is not as good as those schemes which we have already dismissed as being inadequate.

5.6.5 Conclusions

Obviously, we expect any quality control scheme which we use, to signal an out of control situation when the measuring process is indeed out of control. But just as importantly we must not stop the process if it is in control. That is we must not stop the process too often through false alarms. This topic is connected with the operating characteristic of the quality control scheme. This in turn is related to a statistical term called the power of the test or scheme. Saracci (1974) discusses this topic giving medical examples.

One important point made by Hoffmann and Waid (1963) should be borne in mind when considering the last three methods.

'Laboratory personnel know in advance that a reference sample is being tested and they know what values should be obtained from the test. The reference sample was not obtained from a venepuncture, nor was an anticoagulant added to it. It was not transported from the floor, nor was it subjected to the numerous minor factors that can influence results when specimens from patients are handled'.

The medical implications of quality control programs are discussed by Whitby (1969). In my opinion, a standard quality control scheme for all laboratories cannot be constructed. However, no matter what scheme(s) is employed, some basic principles must be exhibited. I think it would be obvious to most practitioners that laboratory size

and the laboratory personnel will affect any quality control scheme employed and hence any scheme implemented must take these factors into account. The main aim of a quality control scheme is to signal out of control situations as soon as possible without stopping the laboratory's work too often for false alarms. We must have, in some sense, a minimal-maximal scheme. That is a scheme which devotes a minimum amount of resources to actual quality control but gives a maximum belief in the reliability of our measuring process.

For a discussion of quality control in a small laboratory, how it affects personnel and the cost, see Padmore (1969). For a general discussion on the implementation of a quality control scheme, see Mitchell (1969).

The situations which we have discussed above are relatively simple in comparison with the current situation found in laboratories in practice. Single determinations on patient specimens, although common, have given way to multiple determinations on single specimens. Obviously, we must guarantee reliability in these situations just as in the single determination situations.

5.7 Quality control for Multiple Determinations.

We have already discussed the Coulter-S autoanalyser and other similar machines. These machines are capable or reporting seven or more clinical parameters for each patient specimen analysed. Many practitioners construct a multiple quality control scheme for such situations which is merely a collection of 'single' schemes. This as we show has some important ramifications. Basically, such schemes do not do what we hope of them. Further, they substantially increase administrative and other quality control work load. Such plans do not use all the information and certainly do not use information efficiently.

Suppose that we have a quality control scheme which signals a stop of a single determination measuring process 5% of the time when the process is operating properly, then a multiple scheme with r determinations being made on each specimen and acting independently in each channel, will stop $(1 - (0.95)^r)\%$ of the time when each channel is operating properly. For r=10, this means approximately 40% of the time. Of course, this result is heavily dependent on the unrealistic assumption that there is independence between each of the determinations being made on each specimen. However, it illustrates the problem.

Instead of drawing up one or at most two control charts for a single determination situation - one chart to monitor the process mean and the other chart to monitor the process precision - we must set up between r and 2r such charts for the r simultaneous determinations. This may not be difficult if we employ a laboratory computer.

Jackson (1959) has discussed a method of quality control for several related variables. Basically it uses the concepts found in Principle Component Analysis. We can describe it briefly as follows.

Let us denote by $\underset{\sim}{x}' = (x_1, x_2, \ldots, x_n)'$ the vector of the measured clinical parameter for a single patient. If we consider the Coulter-S, we may have x_1 = R.B.C., x_2 = W.B.C., x_3 = Hb., x_4 = M.C.V., x_5 = M.C.H.C., x_6 = MCH, and x_7 = Hct., and n=7. We assume that $\underset{\sim}{x}$ has a multivariate Gaussian distribution; a transformation may be necessary. Let the covariance matrix for $\underset{\sim}{x}$ be S and consider the transformation given by

$$\underset{\sim}{y} = W \underset{\sim}{x} \quad \text{such that } W'SW = I. \text{ (I the nxn identity matrix).}$$

W is the matrix of eigenvectors of S' scaled by the eigenvalues. If the assumption of a multivariate Gaussian distribution holds, then $\underset{\sim}{y}'\underset{\sim}{y}$ has a T^2 distribution. The quality control chart now plots T^2 against the sample number. Practically, we would use a quality control specimen for the specimen which generates the $\underset{\sim}{x}$. The system

$$\underset{\sim}{y} = W \underset{\sim}{x}$$

would be required to 'make sense' and we may well envisage using a matrix W* as an approximation to W to give sensible coefficients. We require W* to have the following properties.

(i). Orthogonality (approximate)

(ii). Yield independent $\underset{\sim}{y}$ (approximately)

(iii). To have a high correlation with W.

We note that since S is an estimate to the population covariance structure, checks should be made periodically on W* and S.

The main problem is that we are using a single index for each channel to monitor the multivariate situation. These indices are the components of $\underset{\sim}{y}$. When an out of control signal has been given, we do not know which of the channels is out of control. Thus, we require to check the individual components of $\underset{\sim}{y}$ (or $\underset{\sim}{x}$) against the target vector $W \underset{\sim}{\mu}_x$ (or $\underset{\sim}{\mu}_x$).

We discussed during the investigation of the quality control of a measuring process for a single clinical parameter, the importance of monitoring the precision of the process. In the scheme above, no monitoring of precision is carried out. In a multivariate situation the covariance structure should be monitored. This is the extension to variance in more than one dimension.

Two different covariance structures may be considered. Firstly

the covariance of replicate results. For the seven channels of the Coulter-S, the covariance should be such that the covariance matrix has diagonal form. That is, the off-diagonal elements should be zero, while the diagonal elements should give the squared precisions of each channel. When the matrix is diagonal, or very nearly so, the channels can be considered to be measuring independently. Further, if the diagonal elements are of the magnitude of the precisions squared, the 'precision' of the process (machine) can be considered to be in control.

The second type of covariance structure which may be investigated, concerns the covariance of the patient results. Fundamentally, this should reflect the population covariance structure, i.e. the covariability of the clinical parameters in the population. To monitor this, we require the measured specimens to be a random sample from the population. As we have already indicated, this will rarely, if ever, be achieved. Thus this type of monitoring must suffer the same fate as the schemes discussed in section 5.6. Hamilton and Davidson (1973) have investigated the stability and the inter-relationships of the measurements made by the Coulter-S using 104 medical inpatients. A further problem relating to the analysis of covariance structures concerns the distributional assumptions which must be made before statistical inferences can be made. The basic assumption is that the random variables concerned should have a joint Gaussian distribution. This assumption appears to be unrealistic in practice.

This type of multivariate control is beyond the scope of the laboratory which does not have a computer back-up. For those laboritories having such computer facilities, the scheme, based on

Jackson's procedure is possible. However, in practice, I doubt whether such a scheme will give better control than using control charts for WBC, RBC, Hb and MCV. (Further, it will almost certainly be less intuitively appealing to laboratory staff). The main reason for this, is that the measurement processes on the Coulter are, for all intents and purposes independent. This is unlike a multivariate inductrial situation in which, for example, we wish to control the quality of ball-bearings being produced by monitoring their weight, dimension etc. In such situations there is an obvious dependency of the measured quantities.

We mention a proposed routine quality control scheme for multi-channel analysers described by Laessig, Schwartz, Paskey and Indriksons (1975). They propose a scheme of six pools of analysed reference sera to be used, in a random manner, on clinical chemistry autoanalysers such as the SMA/ 12/60 (Technicon Instruments Corp., Tarrytown, New York) or the Mark X (Hycel Corp, Houston, Texas). The drawbacks of such a scheme has already been indicated in the above. Basically, even if it were to be introduced the amount of computation is prodigious and can only realistically be undertaken if a laboratory has computing facilities (or a sophisticated programmed desk calculator). Further, it only provides an indication of the control of the device; other quality control procedures require to be operated simultaneously. Finally, the problem of carryover is acknowledged but not discussed, which leads us to conclude that as a practical scheme in haematology, with reference to control of the Coulter-S, it has little to offer.

The monitoring of the individual channels has been suggested by Bull (1975). By doing this, we may be able to identify 'patterns' in the results which are symptomatic of particular faults in the measuring process.

We should note that up to this point, we have considered quality control for accuracy and precision. However, we should pay attention to other characteristics of the measuring process. In the multiple determination situation, the investigation of precision should perhaps give way to an investigation of the covariance structure of the process. In a previous chapter we have discussed the carryover effect experienced with continuous sequential autoanalysers. We shall see that we should also monitor this factor in our quality control scheme to be satisfied with the reliability of our results.

We have dealt with control charts and the estimation of certain characteristics of measuring processes. Now we discuss in particular, the control of a measuring process for peripheral blood counts.

5.8 Outline of a Practical Scheme

5.8.1 Introduction

We have noted that quality control should not be concerned only with the work undertaken within the actual measuring laboratory. The aim of the measuring process is to produce reliable and accurate results reflecting the 'true' medical status of the patient whose specimen is being measured. Many factors have already influenced the blood specimen before it reaches the laboratory. In figure 5.16, we present a schematic diagram of the clinical measuring process in which

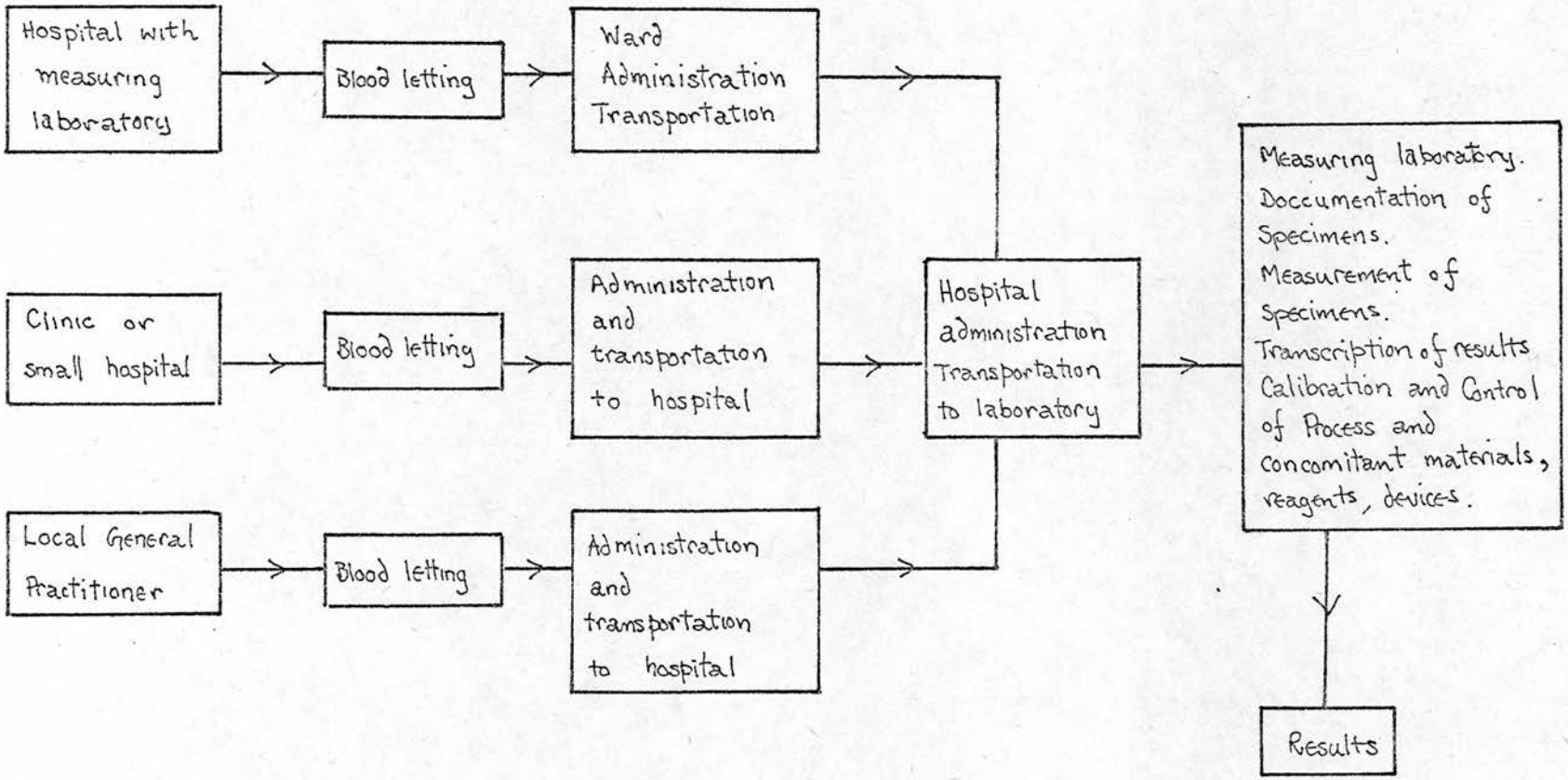

Hospital with measuring laboratory
Blood letting
Ward Administration Transportation
Clinic or small hospital
Blood letting
Administration and transportation to hospital
Local General Practitioner
Blood letting
Administration and transportation to hospital
Hospital administration Transportation to laboratory
Measuring laboratory. Doccumentation of Specimens. Measurement of Specimens. Transcription of results Calibration and Control of Process and concomitant materials, reagents, devices.
Results

Figure 5.1b

we acknowledge the various possible sources of blood tests. A similar diagram was presented in chapter 3, see figure 3.1.

The measuring process as we have noted, describes the complete process outlined in the diagram. In many cases, when authors discuss quality control, they really discuss quality control within the measuring laboratory. Even then, they usually only refer to the mechanistic process within the laboratory, forgetting the biological factors. The diagram does not attempt to show the relative importance of any of the particular parts. However, it is obvious that even with 100% reliability from the laboratory, the actual reliability of the figures on which clinical decisions are made, are also dependent on the reliability of the other parts. Further it cannot exceed the reliability of the most unreliable part, and usually will be much worse.

The collection of blood will commonly be made by student doctors, junior and experienced hospital doctors, general practitioners, and by blood letting teams which may include senior nursing staff. Further the collecting methods and apparatus will almost certainly differ e.g. tourniquet methods, speed of blood letting, time before depositing in collection tube, the bore of the needle, the type of syringe, the anti-coagulant agent, the amount of anticoagulant, the mixing of the blood.

Sacker (1975) has described how the assayed blood will be affected by the specimen collection. The type of blood - capillary, venous and less importantly in this case arterial blood - will yield different results for the same parameter. The collection of capillary blood has to be carefully monitored, since skin temperature, depth of incision, free flowing blood or not and bruising (coagulation), will

all affect the results of subsequent measurements. In the case of venous blood, Sacker describes how the choice of needle, haemolysis due to alcohol and posture will affect blood parameters. Further studies involving the choice of needles, flow rates, containers for specimens and anticoagulant agents have shown that variable results can be obtained if any of these factors are altered.

Chanarin (1975) has reviewed the effects of collection on the Packed Cell Volume (PCV). The PCV increases after a period of recumbency, after tourniquet (dependent on the length of the restriction), after vigorous exercise and after two weeks in bed. The PCV decreases after meals, in ambulant subjects, in the prescence of excess anti-coagulant (not Heparin). Chanarin gives the average increases expected in the other parameters. We note at this stage the difference between the PCV which relates to spun blood, and the Haematocrit which is the calculated counter-part and is presented by Coulter, for example, in place of the PCV.

The administration and transportation of the collected specimens both internal and external - where appropriate - to the hospital will also affect the quality of the results. The actual physical handling of the blood could affect the suitability of the sample e.g. fragile cells. The administration associated with each blood specimen, be it at the clinic or ward or internal to the measuring laboratory is of vital importance. Mismatched samples and report cards can be identified provided gross mistakes are present, but this will not always be so.

It is almost universally accepted that the quality of work within the clinical measuring laboratory should be monitored. However, in some laboratories, the implementation of laboratory

quality control schemes initially met with some opposition, or at least apathy. Most laboratory and clinical staff now realise the importance of such schemes in the management and care of patients. A proposal to introduce quality control in the areas identified above, could produce a similar reaction. With co-operation and explanation, the need for such schemes would hopefully be understood and supported. Major difficulties present themselves due to the variety of origins and originators of the specimens.

Let us consider the situation where bloods are collected within the hospital in which the measuring laboratory is situated. All staff responsible for bleeding patients have received at some time, information relating to the collection, methods and apparatus, of blood specimens. This will have been gleaned from a variety of sources probably over a period of time. Although within a single hospital the apparatus and reagents will be uniform, their usage and the method of collection will vary. It seems almost impossible to suggest how quality control may be carried out on the actual blood-letting. Perhaps the only way in which it may be monitored is by staff vigilance. This normally will be satisfactory, but under heavy work loads or emergencies, discrepancies could result. The best hope is for the staff concerned to realise the importance of following the correct (specified) procedures.

The ward administration is yet another difficult area to monitor. Some authors have advocated that a member of the laboratory staff arranges that specimen(s) are passed through the system, blind to all staff. The specimens can be identified at the results stage. Obviously, in all cases, the full co-operation of staff - medical and nursing - must be obtained. If such proposals are considered

for implementation, the decision will be taken in the light of the workload on haematology staff, nursing staff and other medical personnel. Further, the system must operate in an objective and well documented manner.

The effects of such proposals will be twofold. Firstly, an accurate estimation of administrative errors can be obtained, and secondly, with more uniform collection methods etc. less variability will be experienced. The effects of this on the clinical 'normal' ranges are obvious.

We are predominantly concerned with peripheral blood counting, although much of what has been said is pertinant to other clinical investigations. Let us consider the actual operation of quality control within the laboratory. We have two areas with which to contend, namely the mechanistic and the biological factors which we have previously mentioned. As an aid to the discussion, we give a description of a scheme currently being used in the Haematology Department of the Edinburgh Royal Infirmary. It must be made clear at the outset that the scheme has been evolved and indeed is still evolving. The idea that a unique quality control scheme can be drawn up for all laboratories is quite wrong as most haematologists and technicians will appreciate. Quality control is dependent on the laboratory size, type of work, administration, laboratory staff, machinery, etc. However, we emphasise that we believe that certain basic principles must be adopted. The scheme described has been employed in conjunction with the Coulter-S, and we shall give examples with this in mind.

5.8.2 Reference Preparations

The commercial preparations used in the scheme is the manufacturer's own - 4C, Haematology Reference Control for Coulter

Counter Model 'S' Automated haematology System, manufactured by Coulter Electronics Limited, Hertfordshire. Another commercial control had been used but has been removed from the market; this will be replaced in the scheme when a suitable material can be found. The '4C' is used to calibrate the Hb, RBC, WBC, MCV, MCH and MCHC channels. The MCV channel is continually monitored using 'walking controls'. A pool of approximately 40 normal donors is used. (The Mean Cell Volume has been found to be remarkably stable).

Any new batch of 4C arriving in the laboratory is tested against the previous batch and the results are compared with the values supplied by Coulter on the accompanying assay sheet. A batch of 4C consists of eight phials, which are numbered by the laboratory 1, 2, 3, ..., 8 - each bottle can be sampled six times only. The first bottle is kept separate for checking purposes. On arrival it is sampled three times, and then stored at $4^{o}C$ until the next batch arrives, that being when this present batch is almost exhausted. The remains of this first bottle are compared in the same way with the first bottle of the new incoming batch.

The results of the three sampled specimens from bottle 1 are used in the following way. The first is discarded, the second and third, in order, being averaged. Two comparisons are made. The mean of the new batch is compared with the Coulter Assay values. The mean of the previous batch - the final samples of the old bottle 1 - is compared with its first assay values i.e. the values obtained when it first came into the laboratory.

Let us consider an example of the comparisons which are made. The data are given in table 5.8. (We note that changes in reagent status is assumed to have been monitored and accepted).

New Batch			Previous Batch	
Lab. results	Assay sheet		Final results	Initial results
9.2	9.0	WBC($10^9/_1$)	9.2	9.1
4.96	4.95	RBC($10^{12}/_1$)	4.97	5.00
14.6	14.6	Hb.(g/dl)	14.5	14.5
0.436	0.426	Hct.	0.431	0.435
89	86	MCV(fl)	88	87.5
29.4	29.5	MCH(pg)	29.6	28.8
33.5	34.3	MCHC(g/dl)	34.2	33.3

TABLE 5.8

The comparison of 'previous batch' allows investigation of the consistency of the laboratory. Thus when faced with a result, for example MCV of 89 fl. as measured in the laboratory and a Coulter value of 86 fl. we can be reasonably confident - also due to the other components of the quality control scheme which we shall discuss - that we should point out the discrepancy to the manufacturer. In fact, in this case, the manufacturer subsequently sent out a letter pointing out an increase in the MCV for this batch.

The testing of 4C is carried out at the machine set-up stage to enable differentiating between machine drift and changes between batches.

Reference Blood

Donor blood is used for day to day control. A batch of such blood usually lasts for about two weeks. Let us consider three batches of such control (donor) blood. The batches are taken at weekly intervals into the quality control scheme. We represent the usage of the

three batches schematically in figure 5.17.

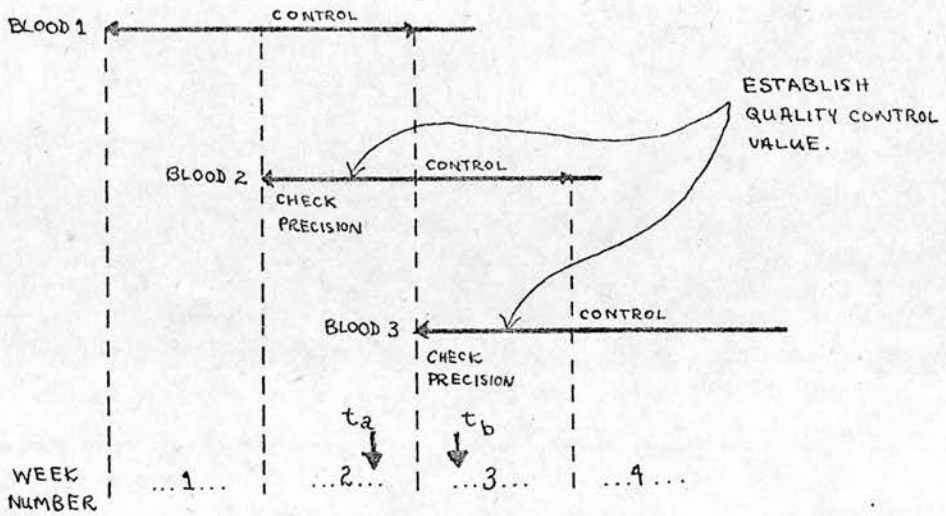


FIGURE 5.17

By using this overlapping reference system, we have a reasonably proficient 'fail-safe' system. Consider blood 2; if at time t_a, when blood 1 is being used as the control blood, we find serious accuracy and/or precision problems, we have as well as the commercial preparation, blood 2 to compare against. Even at time t_b, there will almost certainly be blood 1 in sufficient quantity to use as a second control blood to blood 2.

Before discussing the day-to-day control of the Coulter-S measuring process, we emphasise the need for the principle measuring device to have undergone a full evaluation (see Chapters 3 and 4) and also to have undergone a period of extensive control and comparison at its introduction to the measuring laboratory. Having made this point, we now describe the day-to-day control which is practised in the laboratory.

5.8.3 The Daily Set-Up

At the beginning of each working day the Coulter-S is 'set-up' using the following procedure. The machine is fully cleaned, bleached and rinsed. Background counts are obtained by averaging the results of replicate counts on diluent and Lysing agent. The maximum acceptable values for these counts are $\pm$ 0.0 for Hb, 0.02 for both RBC and WBC. Throughout this period the electronics and the pneumatics are checked. For example, the flow rates are inspected and the presence of bubbles in the tubes are controlled. Having carried out this, the following sequence of blood counts are made.

(i). Three specimens of 4C are counted sequentially. The results of the first specimen are discarded (effects of carryover). The results for all clinical parameters for the second and third specimens are compared with the established values for this batch of 4C. These established values, which are the laboratory's own, may differ from the Coulter values. (We have seen in section 5.8.7 that differences can occur and this may be due to an inaccuracy of the Coulter value). Further the tolerances used by the laboratory are often stricter than those suggested by Coulter. Given that the Coulter values for 4C are checked on arrival of the batch in the laboratory, see 5.8.2, the use of this check is particularly useful in monitoring WBC and MCV.

(ii). One of the laboratory's own quality control bloods, 'normal', is counted twice sequentially. The first set of results are discarded. The second set of results are

compared with the laboratory's established values for this blood.

(iii). Another quality control blood which has been established by the laboratory is counted twice sequentially. This blood has reduced WBC, RBC and Hb. Again the first set of results are discarded while the second are compared with the established values. This, together with the check detailed in (ii) monitors accuracy at two points. Unfortunately, it is usually not possible to obtain a blood for such use which also has a reduced MCV value. The MCV is commonly the same value as the blood used in (ii).

(iv). A patient specimen from the previous day, which has a WBC of approximately 10.0 x 10^9/l is counted sequentially twice. The first set of results are discarded, while the second are compared with the results reported for the previous day. These results were obtained at stage (v) on the previous day.

(v). A new patient specimen with WBC of approximately 10.0 x 10^9/l is counted sequentially twice to establish a new blood for stage (iv) on the next day. Further, independent measurements are made on Hb, WBC and PCV. The WBC is counted visually, the Hb is measured twice using a colorimetric method and finally the PCV is determined by centrifuging. We note that it is the PCV which is established by this method, not the Hct which the Coulter-S reports. These values can be discrepant due to trapped plasma. To minimise the possibility of

such discrepancies blood specimens with approximately 'normal' RBC are used.

(vi). Finally, carryover is checked using two quality control bloods, A and B. The first blood A has 'normal' clinical parameters e.g. the blood used in (ii). The second blood is also a quality control blood and has reduced Hb, WBC, and RBC, e.g. the blood used in (iii). The sequence of presentation is,

D D D B B B A A A D D D

where D represents diluent specimens, and A and B represent blood replicates. Carryover is determined, see 3.5.3. The maximum allowable carryover is 3%. Unfortunately, the carryover for the MCV channel is difficult to check, since as we have indicated previously it is usual that the bloods obtained for use as quality control agents have almost the same MCV.

If at any of these six stages there are unfavourable comparisons (inconsistencies), the set-up procedure is halted until the fault has been found and corrected.

5.8.4 Through the Day Control

The quality control performed by the laboratory throughout the day has several components. We shall call the different facets of this control by the following names.

(i). Routine Quality Control specimens

(ii). Blind Checks on routine quality control specimens.

(iii). Repeat analyses on patient specimens.

(i). Routine Quality Control Specimens

By routine quality control specimens, we mean specimens of the

laboratory's own quality control blood. The procedure for establishing the clinical values for this preparation has been discussed in section 5.8.2. Two control specimens, more if necessary, are counted sequentially after every thirty specimens. (This can be changed to every twenty, if serious control problems have been experienced). The interval between these control specimens depends on the amount of control blood and also on the current day's workload.

We note that each control specimen requires 1.4 ml of blood, and so usually after each batch of thirty routine patient determinations, 2.8 ml of control blood has been used. The 'wastage rate' on such blood is substantial - that is if we consider that blood used in this way is wasted.

The Coulter operators know when these specimens are presented and further they also know the 'established' values i.e. the values which should be obtained if the process were operating correctly. We have stated previously that in such circumstances the operators do take more care over these specimens. These specimens are seen as a 'confidence' check for the operators. That is, they allow the operators to assess the status of the machine. In other words, this is not in essence an 'operator check', it is a machine check. As we see in the following discussion, replicate and blind quality control specimens check the total laboratory measuring process.

Represent routine patient specimens by X and the quality control specimens by S. The usual sequence of presentation to the machine is given in figure 5.18.

$$\ldots X^{*}\ S_{11}S_{12}\ \overbrace{X\ X\ldots X\ X^{*}}^{30}\ S_{21}\ S_{22}\ \overbrace{X\ X\ X\ \ldots\ X^{*}}^{30}\ S_{31}^{\uparrow}S_{32}^{\uparrow}\ \textcircled{or}\ \begin{matrix} S_{33}^{\uparrow} \ldots \text{check on process} \\ S_{33} \ldots \text{routine use.} \end{matrix}$$

FIGURE 5.18

The values of the patient specimens which have been asterisked are noted. The specimens S_{ji} are the replicate control blood specimens. At least two control specimens are measured. If an out of control signal, i.e. an indication of a fault, is given e.g. at $S_{31}^{\uparrow}$ and $S_{32}^{\uparrow}$ a third control specimen is immediately counted. (The arrow denotes out of control). Two results are possible, as shown in the figure 5.18. We could obtain S_{33} whence the out of control $S_{31}^{\uparrow}$ and $S_{32}^{\uparrow}$ are attributed to chance. Alternatively we could obtain $S_{33}^{\uparrow}$, in which case the third control specimen also gives an out of control signal. In this case the process is stopped and the cause of the 'out of control' is investigated. If the cause is detected as trivial, quality control specimens are counted after its correction to ensure the correction. If the cause is detected as major, a full and complete calibration check, as at the set-up for the beginning of the day, is undertaken.

From these replicate quality control specimens information concerning the accuracy of the machine is obtained. It may also be possible to estimate and monitor carryover. To explain how this is done we discuss each of these in turn.

(a) Information on accuracy

It will be obvious that the first of the control replicates S_{j1}, $j = 1,2,\ldots$ will often be affected by carryover from the preceding patient specimen. Since we cannot control which patient specimens come before the control replicates, the effect of

carryover will be variable. For this reason the first of the replicate specimens results are discarded. The second set of results, i.e. those relating to S_{j2}, $j = 1,2,\ldots$ are used to monitor the accuracy of the machine. (We note that this as we have stated above is not necessarily an objective method of control).

The values, S_{j2} are compared with the established values of the quality control blood. A range of values for each parameter, are set-up with the mid-point as X the established value. The width of the intervals obviously depend on the precision of the specific channel, but experience also does usually have some influence. If the quality control specimen values are within the ranges, the process continues. If the counts are without the ranges a third quality control specimen is immediately counted. We have stated the procedure followed with these results in the above.

Is this a reasonable control procedure? To answer this question we discuss it in terms of the mathematical model proposed in this section. That is,

$$y_t = \alpha_o + \alpha_1(1-\theta)X_t + \alpha_1\theta X_{t-1} + dt + \varepsilon_t$$

where y_t is the t^{th} measurement on a specific channel, X_t, X_{t-1} are the true values of the t^{th} and $t-1^{th}$ measurement. θ is the carryover, d is the drift, α_1 denotes the linearity, α_o denotes the calibration and ε_t is a random error associated with the t^{th} measurement.

Let the 'true' value of the control bloods be μ. Then we have for example

$S_{12} = \alpha_o + \alpha_1(1-\theta)\mu + \alpha_1\theta\mu + dt + \varepsilon_{12} = \alpha_o + \alpha_1\mu + dt + \varepsilon_{12}$

Assume that the error ε_t has a Gaussian distribution with zero mean and variance σ^2. Then S_{12}, in fact all second control replicates have a Gaussian distribution with mean $\alpha_o + \alpha_1\mu + dt$ and variance σ^2. When the process is in control we have $\alpha_o = 0$, $\alpha_1 = 1$ and $d = 0$. Let us assume that the control range is $\mu \pm L$. Then the probability that an out of control signal is given when the process is in control is $1 - 2\Phi(-\frac{L}{\sigma})$. Thus, for example, when $L = 2$, 4.55% of the time when the process is in control it will be signalled as out of control. Of course the use of a third quality control specimen in such cases leads to fewer false alarms. In fact these comments are similar to comments which were made when discussing modified Shewhart charts.

Thus not only does this check provide the operators with a 'yardstick' with which to monitor their own performance, it also gives a reasonable indication of a lapse in accuracy if it occurs. Further, due to the frequency, i.e. every 30, this means approximately 700/30 replicates per day. That is there are approximately 25 of these replicate quality control specimens per day. Thus there is a reasonable amount of data accumulated with which to investigate the process performance. When the process is in control this accumulated data will be randomly distributed about the established quality control value. Any specific bias is indicative of a loss of control.

(b) Information on Carryover

The main problem with trying to obtain an indication of the carryover from the replicate control specimens is that we require

to know the preceding patient specimen values. However, even if these are known, frequently the values are so close to the quality control established values that carryover cannot be estimated.

When a difference in the clinical values of the control and patient specimens exists, carryover is monitored in the following way. A rough indication of the carryover status of the channels is given by considering the difference between the two control replicate values. Carryover is deemed to be acceptable if the difference is no larger than 3% of the established quality control value for the control specimens. However this information cannot be used in isolation. If, for example the preceding patient specimen had severely abnormal values the difference may well be in excess of 3% of the established values, although not caused by a carryover fault. Thus even with this simplified method of monitoring carryover status, the preceding patient values must be consulted on occasion.

In chapter 3 we discussed the definition of carryover and also methods of estimation. Following from this it may be thought that we could employ such methods in this situation.

When a difference in clinical parameters exists between control specimen and patient specimen, we could estimate carryover K, where e.g.

$$K = \frac{S_{11} - S_{12}}{X - S_{12}}$$

Assuming that the drift between successive measurements is zero and that measurement errors are small, we have seen in chapter 3

that the expected value of K is approximately θ. Thus K gives an indication of the carryover present in the current process. (We note that the established value of the quality control specimen should not be used in place of the second specimen value. If this were to be done the resulting estimate of carryover would be affected by any change in accuracy/linearity/drift since the control values had been established. We remark at this stage that if an index is constructed to monitor a specific process parameter, it should not be affected by other process parameters).

To see the probabilistic interpretation which may be given to the index K we consider the following. It can be shown, see section 3.10, that if μ is the true value of the control blood, μ_x is the true value of the preceding patient specimen, θ is the current carryover, σ is the precision of the channel and σ_x is the biological variability of the channel, that the probability that K exceed the value λ is given by:-

$$P(K>\lambda) = P\left(z > \frac{-\alpha_1(\theta-\lambda)(\mu_x-\mu)}{\sqrt{2\sigma^2(1 - 2\lambda + 2\lambda^2) + \lambda^2\sigma_x^2}}\right)$$

where z has a standard Gaussian distribution.

Consider the following numerical examples in table 5.9, which relate to situations which could arise for the Hb channel.

TABLE 5.9 on next page

	$\mu_x-\mu=2.0$ $\sigma^2=(0.1)^2$ $\alpha_1=1$ $\sigma_x^2=(2.0)^2$ $\theta=3.0\%$	$\mu_x-\mu=2.0$ $\sigma^2=(0.1)^2$ $\alpha_1=1$ $\sigma_x^2=(2.0)^2$ $\theta=2.0\%$	$\mu_x-\mu=1.0$ $\sigma^2=(0.1)^2$ $\alpha_1=1$ $\sigma_x^2=(2.0)^2$ $\theta=3.0\%$	$\mu_x-\mu=1.0$ $\sigma^2=(0.1)^2$ $\alpha_1=1$ $\sigma_x^2=(2.0)^2$ $\theta=2\%$
Probability that K exceeds λ				
$\lambda=3$	0.500	0.448	0.500	0.472
$\lambda=4$	0.448	0.401	0.476	0.448
$\lambda=6$	0.413	0.386	0.456	0.440

TABLE 5.9

We note that even when carryover is 2% we expect there to be a large percentage of times in which the value of K obtained will exceed the permissible value of 3%. Further we see that the smaller the difference between the control value and the preceding patient specimen value the larger the frequency of false alarms. The essentially commonsense way in which the laboratory handles this carryover information is reasonable and does not acknowledge that there is more than an indication of the status of carryover present in the figures. The more sophisticated probabilistic approach is not suggested for use.

(ii). Blind checks on quality control specimens

In (i) we noted that not only did the operators know when the replicate quality control specimens were counted, but they also knew

the values which should be obtained. We note that this gives the operators a 'yardstick' with which to assess their performance. A more objective measure of the reliability of the process is possible through the use of 'blind' checks on quality control specimens. 'Blind' describes the fact that the operators do not know when these checks are made. In other words, the 'blind' controls are presented to the measuring process not known to be control specimens, at least not until the results are obtained. In practice single specimens of the routine quality control blood are presented for counting. Only the chief technician/quality control officer knows which specimens constitute the 'blind' checks. The check specimens are made on average every 60^{th} specimen, although obviously at random. The specimens' results give a more objective indication of the quality of the results.

Let us denote the random 'blind' check specimens by B_i, $i = 1, 2, \ldots$ Since the blind checks are randomly presented, the patient specimen preceding the control specimen could be low abnormal, normal or high abnormal. Due to carryover the effect of the preceding blood on the control specimen cannot be ignored. By using the model given in section 3.7, we investigate the use of these 'blind' checks. Let the 'true' value of the control specimen be μ and let the true value of the specimen preceding the control blood be μ_{x_i}, $i = 1, 2\ldots$. Then we have the following representation for B_i,

$$B_i = \alpha_o + \alpha_1\mu + \alpha_1\theta(\mu_{x_i} - \mu) + f(t_i) + \varepsilon_i$$

where $f(t_i)$ is the drift at the time of measuring the control blood. Further, the control value of the quality control specimen is B, where

$$B = \alpha_o + \alpha_1 \mu$$

Commonly we will assume that $\alpha_o = 0$ and $\alpha_1 = 1$. We construct the differences $B_i - B$. From the above, we have

$$B_i - B = \alpha_1 \theta(\mu_{x_i} - \mu) + f(t_i) + \varepsilon_i \quad i = 1, 2, 3, \ldots\ldots$$

The true value of the preceding blood μ_{x_i}, is a random variable with respect to the differences $B_i - B$.

A sophisticated statistical analysis concerning the $B_i - B$, $i = 1, 2, \ldots$ is not appropriate, due to the number of influences affecting them. However, the following graphical display, which is carried out in the laboratory, can give useful information.

The successive differences $B_i - B$, $i = 1, 2, \ldots$ are plotted against i, which in this situation is simply time, see figure 5.19.

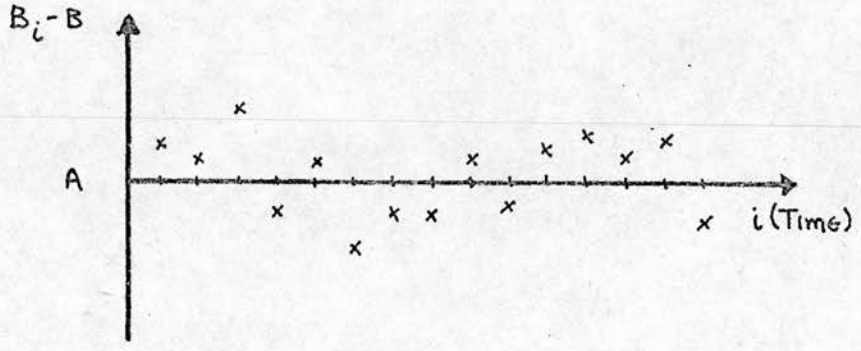


FIGURE 5.19

Against the vertical axis the difference between the current measured value and the established value is plotted. If the quality control blood has 'normal' values, the horizontal axis occurs with A approximately zero, when the process is in control.

By inspecting the graph, we may be able to detect persistant biases. That is particular patterns may occur. For example, see figure 5.20.

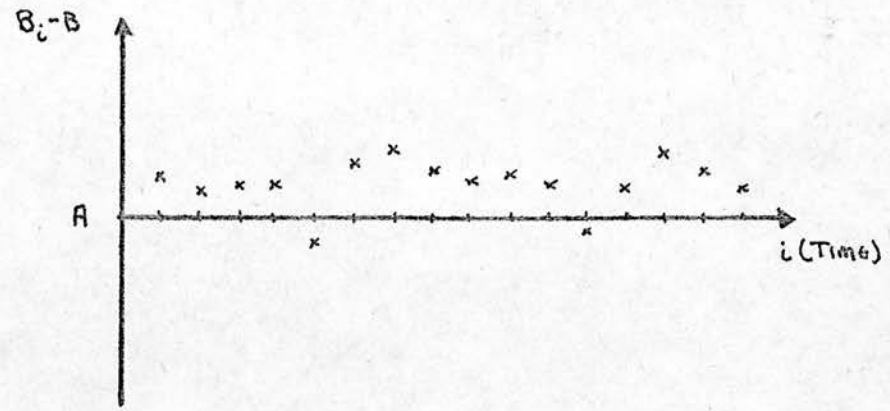


FIGURE 5.20.

It is unlikely that the pattern in figure 5.20 has been caused by the 'random' chance that a 'high' blood specimen preceded every control specimen. Thus, we conclude that the process accuracy has changed. (Changes in linearity and/or carryover would not produce a consistent pattern such as this).

When the process is in control a pattern such as that in figure 5.19 will be seen. That is there should be a random scatter about the horizontal axis.

(iii). Repeat analyses on patient specimens.

The control afforded by the procedures described in (i) and (ii) depend on quality control blood. This has two major disadvantages. Firstly, the quality control blood monitors the process at a single clinical value - in each parameter. This value is nearly always in the middle of the clinical 'normal' range. Secondly, the quality control blood has been 'preserved', i.e. it is not a fresh sample, Although the use of quality control blood is important in monitoring the process, it does not monitor the handling of freshly sequestrened patient blood specimens and their storage.

To overcome these problems, selected patient specimens are also used in controlling. (Further, the results of these specimens reflect the quality of the laboratory's work, from logging-in of specimens, to measuring to the transcription of patient results, as do the blind quality control specimens).

Each day, residual blood, after the specimen has been counted routinely, from 20 patient specimens covering the clinical ranges in the parameters are set aside. On the following day, these specimens are counted at random intervals throughout the working day.

Let R_i and R_i' denote the measurements made on the i^{th} patient specimen on the first day and second day respectively. Let ξ_i be the true value of the i^{th} blood and let μ_i and μ_i' be the true values of the preceding blood on the first day and second day respectively. As before, we can write for the first day's result,

$$R_i = \alpha_o + \alpha_1(1-\theta)\xi_i + \alpha_i\theta\mu_i + f(t_i) + \varepsilon_i$$

On the second day, the parameters are denoted with dashes to indicate a possible difference from the preceding day. Then we have

$$R_i' = \alpha_o' + \alpha_1'(1-\theta')\xi_i + \alpha_1'\theta'\mu_i' + f'(t_i') + \varepsilon_i'$$

The difference between the readings on the same specimen on different days is computed, i.e. $R_i' - R_i$. This difference can be written as:-

$$R_i' - R_i = (\alpha_o' - \alpha_o) + \{\alpha_1'(1-\theta') - \alpha_1(1-\theta)\}\xi_i + \alpha_1'\theta'\mu_i' - \alpha_1\theta\mu_i + f'(t_i) -$$

$$- f(t_i) + \varepsilon_i' - \varepsilon_i$$

The following sources of variation are present in these differences

- difference between day to day calibration
- difference in carryover
- difference in drift
- difference in preceding patient specimens
- natural process variation i.e. random error.

The differences $R_1' - R_i$ look to be so affected by a multitude of factors, that it is difficult to see how we may use them. However, the other components of quality control furnish information concerning carryover and accuracy (drift). In the light of such information for the first and second days, some information can be gained from the differences $R_i' - R_i$. The method used for inspecting the data is similar to that described in (ii). The successive differences $R_i' - R_i$ are plotted on a graph which has a horizontal axis at zero, see figure 5.21.

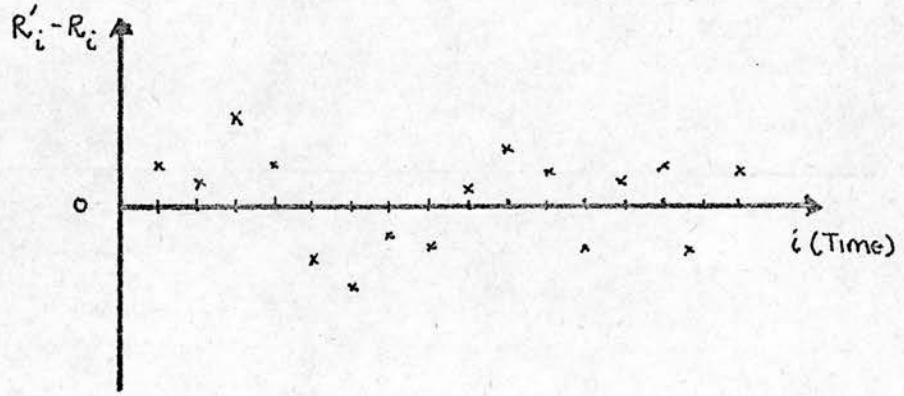


FIGURE 5.21

Once again any discernable and persistent pattern is indicative of a control problem. To ascertain the cause, the quality control information must be viewed in total.

5.8.5 Comments on the Scheme.

We have discussed a scheme of intra-laboratory quality control which is currently in use. We have already indicated that in practice, quality control schemes will evolve, and this is certainly true of the scheme which has been described. The evolution, from some basic principles, has two benefits. It not only allows development of the scheme together with the 'education' of the technical and medical staff, it also allows for new problem areas of control to be identified. Once these areas have been identified, procedures for control can be introduced. A further benefit, is that the particular laboratory can create a scheme which 'fits' its own specific requirements. This brings the discussion round to a consideration of the 'cost effectiveness to current medical practice' of the quality control scheme which is practised. Improvements to the scheme described, e.g. the analysis of more control specimens, the use of laboratory computers, etc., are costly in terms of money and also in an increase of the workload of the laboratory. For a quality control scheme to 'fit' a specific laboratory it must ensure the reliability of results needed by the medical staff for as high a standard of medical care as can be afforded.

As we have stated the quality control scheme which has been described, has evolved. This evolution results in a reasonable scheme for a laboratory which handles approximately 700 peripheral blood counts per day. The laboratory does not use a laboratory or other computer. (The use of nearby computing facilities would result in a loss of simultaneous and immediate control). The computation involved in the quality control scheme described is simple and can be performed quickly, if necessary by any member of the technical staff. As we

have indicated previously, this is a desirable property. It means that the quality control is not remote from the laboratory staff. Further, the procedures which have been described are intuitively reasonable. Thus the staff will, in practice tend to believe the indications. This may not happen in situations in which sophisticated indices and arithmetic are used.

The intra-laboratory quality control scheme is the sum of all the individual components which we have described. Each part of the scheme is important in its own right. But further, this is definitely a situation in which the sum of the parts is greater than the whole.

As we have seen the scheme allows for the accuracy (drift) and carryover to be monitored. We note that little has been said concerning the monitoring of the process precision. In fact the precision of each channel is monitored - primarily at the beginning of each day and if necessary throughout the day. It is monitored using replicate counts on the quality control blood. Some information is present from the accumulated quality control data on the second of the replicate measurements described in (i). If these measurements are wildly dispersed with no other assignable cause, a check on the process precision is warranted.

Even without the aid of a computer, the laboratory monitors all the facets of the process. For example the use of patient speciments can yield some information concerning the linearity of the process. It will be obvious that information relating to the exact status of the process cannot be viewed in isolation. All quality control data must be used simultaneously.

It is impossible to construct a quality control scheme which identifies the cause of the loss of control. To do this would require

a scheme capable of identifying the effect of every single factor on the process. However with experience, certain patterns in the loss of control can suggest specific causes. To see how varied the causes of loss of control can be, we give the following example.

Whatever we wish or hope for, the Coulter-S is simply a particle counter. Although we wish to count cells, that may or may not be what we are actually achieving. At a time in the laboratory, with new operators, starch powder was used to assist the donning of the protective gloves. An increase in cell counts resulted - much as is expected in high pollen count days. The loss of accuracy was detected, but the cause was traced only after the 'usual' loss of control causes had been eliminated. Thus although it is important to maintain a constant vigil on the process, it is as important to be able to analyse the process to find the cause of the loss of control.

As we have discussed in the previous sections, certain quality control procedures which have been proposed in the literature, are doubtful. Among other such procedures is the comparison of the mean of a group of patient specimens at the end of one day with the mean of the same group of specimens on the next day to check the calibration of the machine. The problem with reliance on this method is that if the means show some discrepancy, this is indicative of a change in the process status. However, there is no information as to what has caused this loss in control. More distrubing, is the fact that even when the means show no discrepency, this may be due to a genuine situation of no change in status of the machine. However, it may also be caused by changes in drift, accuracy, carryover, precision and/or linearity combining in such a way to cancel each other out, giving no discrepancies in the means.

Unlike this, the procedures described, i.e. (i), (ii) and (iii) and the other quality control procedures, allow for the identification of the facet(s) of the process which are out of control. This is not to say that there are not problems in using the procedures which we have described. For example, in the use of repeated analysis of patient specimens, which we have discussed in section 5.8.4(iii). In this case, the effect of the preceding 'unknown' patient specimen on the control specimens cannot be easily estimated. However, as we have indicated, a liberal allocation of common sense and experience should enable even these results to be meaningful. Carstairs, Peters and Kuzin (1977) have described the use of 'random duplicates', i.e. randomly selected, current day patient specimens in laboratory quality control. The second of these values, obtained after an interval of time, is accepted if Hb is within $\pm$ 0.4 g/dl of the first result, WBC is within $\pm$ 0.08 x 10 g/dl of the first result. As the authors indicate, these intervals represent approximately 4 standard deviations of the respective channels. In practice, a more flexible scheme such as described above allows for action in conjunction with other quality control information. A rigid scheme such as Carstairs, Peters and Kuzin describe, may well cause an out of control signal to be given each time a pathological specimen precedes either or both of the quality control specimens.

In much of the interpretation of quality control data, a large reliance is placed on experience; not only experience with the particular machine involved, but also experience in the general field of haematology. Further, the technical staff must be aware, and acknowledge the requirements of the clinical staff with respect to the clinical measurements.

Due to the relative cheapness of micro-chips and micro-processors etc., it is probable that a new generation of laboratory machines will become available. (Such developments have already occurred in clinical chemistry e.g. The Technician SMAC). These new generation machines will have automatic, built-in quality control facilities. Even so, the experience of the existing staff will be required to monitor the monitoring of the machines. Further, if such devices are to be marketed, the programs for control and other control devices will have to be created. This can only be done realistically with the help of the current haematology technicians.

6 Inter-Laboratory Quality Control

6.1 Introduction.

In chapter 5, we discussed quality control schemes designed to monitor the reliability of a laboratory by the laboratory itself, i.e. Intra-laboratory quality control schemes. Although most laboratories take care to produce results compatible with commercial standard/reference preparations, it is known that laboratories produce different results with the same or similar methods and machines. The causes of such differences are many and varied, being both chemical and physical. Calibration problems with machines and laboratory equipment, reagent standardisation, deviations from prescribed procedures etc., all contribute to such differences. In this chapter, we consider inter-laboratory control. The relationship between the investigations of chapter 5 and chapter 6 is similar to the relationship between chapter 4 and chapter 4. In chapter 3 we discussed the evaluation of a measuring process in 'isolation'. In chapter 4, we discussed the comparison of a measuring method with other equivalent measuring processes. The similarity lies in the fact that having considered methods for the monitoring of a laboratory's results in 'isolation', now we compare the laboratory's results with those of other laboratories. The size of such schemes can vary greatly from the local, regional to the national and international. Although with increasing size practical problems develop, especially with specimen preservation, the general mathematical treatment is similar.

6.2 Aims of Inter-laboratory Schemes.

Some recurring phrases used in describing such schemes are 'proficiency assessment schemes' and determinations of 'the state of the art'. Neither of these emphasise the usefulness of the collected information to the participating laboratories. Admittedly, the information is often so 'out of date' that it cannot be used in the laboratory's own quality control scheme.

Basically, as Van Assendeft and Holtz (1975) state, the schemes provide investigation in two areas. They can be considered as 'Quality Studies' and/or "Quality Control Studies', not necessarily in that order.

A Quality Study investigates the methods being used in the haematological determination. The study should bring to light information on the robustness of the method, current usage, ease of usage etc. For example when the Coulter-S was introduced, such studies monitored the relative behaviour and developing expertise of the technical staff.

A Quality Control Study is an attempt to control or regulate participating laboratories. Such studies, to be useful to the participants must return information quickly. Local schemes are probably more capable of this than larger national schemes.

In a brief introduction to a collection of papers dealing with inter-laboratory quality control, admittedly not dealing with medical laboratories, Ku(24) has suggested that such inter-laboratory schemes have three purposes.

(i). Trouble shooting, or audit of the comparability of measurements.

(ii). Evaluation of Test methods.

(iii). Extension of a measurement process from a primary laboratory to other standard laboratories.

We shall return to these points throughout our future discussions. However, we see that point (i) relates to the controlling of the measuring processes. Point (ii) relates to the quality study as defined by Van Assendeft and Holtz above. The descriptions 'primary' and 'standard' laboratories are not commonplace in haematology. However as we shall see in the following discussions, there may well be a parallel if we consider research and development laboratories and those, perhaps smaller routine, less well equipped laboratories.

A criticism of certain inter-laboratory quality control schemes, is that the participating laboratories are somewhat bewildered by the feedback from the schemes. This must be considered to be alien to the whole concept of such schemes. The results from the schemes must be meaningful, and potentially useful to the participants. If this is not so, effects of disinterest, lack of understanding, apathy etc, could seriously influence a laboratory's participation, and so affect the overall results from the scheme. Bewilderment at results may reflect a desire to obtain information from the scheme which it was not intended to give.

Youden (1959) perhaps oversimplifies the situation, but he laid great importance on the feedback to the participating laboratories.

'When all is said and done, what we want is rather simple. We want to know whether the test procedure as set forth is capable of yielding acceptable agreement among results from different laboratories. If the results are not acceptable, we would like some specific indication of what is wrong with the procedure. If the procedure

appears to be reasonably good but there are some disturbing discrepancies, we would like to know which laboratories are having trouble. And most important, we should be able to get this information back to the laboratories concerned in such a form that the diagnosis is believed. For only then will these laboratories take any action to correct the difficulties.'

Such inter-laboratory schemes appear to be very distant from the actual patient-care-diagnosis problem which faces the routine haematology laboratory. However, these 'grass-roots' problems are the motivations for such schemes. We must not lose sight of the fact that we are essentially trying to improve and/or maintain the reliability and validity of laboratory results and not just trying to perfect some abstract measuring process, 'or collect information to enable yet another publication to be produced'. This last quotation was taken from a response to the survey described in section 6.9.

We point out that, as yet, participation in such schemes is not a legal requirement. However, it does not appear that such a legal requirement is beyond implementation. This has been the subject of editorial comment in the British Medical Journal, (May 1977, p.1180).

Leaving aside any legal considerations, we can identify two main reasons for participating in such schemes. Firstly, as we shall show, even a comprehensive internal quality control scheme (i.e. a good intra-laboratory quality control scheme), cannot detect certain errors in laboratory practice, e.g. bias, systematic errors, scale errors etc. Participation in such schemes leads to improved haematological practice. Secondly, advisory or standardising committees can use the information collected to suggest new and better methods, materials, etc.

We can be purely altruistic and aim to perfect the 'art'. Howeyer, the practical ramifications cannot be overlooked. The mobility of patients has greatly increased and consistent and transferrable health records should be available. The use of 'normal ranges' in everyday medical practice must be based on consistent and accurate 'scales'.

Some laboratories may regard such schemes as snooping or at least a waste of time. In some cases where laboratories already have a heavy workload, it may be thought to be less important than quickly dealing with patient specimens - even if the measuring process is not compatible with other measuring processes for the same clinical parameter. Fortunately, it seems that such attitudes are dying and most laboratories accept and even welcome these schemes, see section 6.9.

In section 6.9 we describe the DHSS(LDAG) and BCSH haematology quality control scheme. In fact by Van Assendeft and Holtz's description, this is really a 'quality study'. The trials have evolved in both size, content and intent since their inception in 1968. The first trials, (Lewis and Burgess (1969)) although designed to give information regarding differences between laboratories, were primarily designed to give information concerning the materials employed. The emphasis six years later, Ward and Lewis (1975), appears to have moved to cumulative assessments of laboratories and other 'control' information. The effects of this have already been seen, section 5.6.4. Pragnell and Johnston (1977) assess their intra-laboratory quality control by means of the DHSS trial results.

Thus although an inter-laboratory scheme may have been operating for some time, the emphasis may have changed. Further, these schemes may have different values for different laboratories.

6.3 Set-up of Inter-laboratory Quality Control Schemes.

Unlike many of the other fields in which inter-laboratory quality control schemes are used, haematology suffers from a problem with reference/standard materials. Lewis and Burgess (1969) and Lewis (1975) have detailed problems related to establishing standard/reference materials in haematology, see section 5.3.2. Lewis proposed some tentative solutions. Such problems are of fundamental importance, but any further discussion here would divert us from the present investigations. However, care should be taken to ensure that any suggested scheme could be implemented with existing materials.

In any scheme, the participating laboratories must be in no doubt as to what is expected of them, both in handling and analysing the materials and in handling the results of these analyses. They must receive comprehensive and comprehensible literature regarding all facets of the scheme. The literature should cover not only the short and long term aims of the scheme but also information concerning methods of measurement, reagents, the way in which results should be reported and an explanation of the format of the reported results.

The participating laboratories must return enough information to enable them to be assigned to particular sub-groups. Thus, method of determination, equipment, time delays etc. are taken into consideration and each participant assigned to a group composed of similar and comparable participants. These groups should represent, homogeneous collections in which comparisons can be made. Comparison between groups will also often be of interest, e.g. between manual and automatic methods.

Much more could be said concerning the administration of the schemes but to a large extent the administration is dependent on the materials circulated to the participants the geographical scatter of the participants and on the content of the schemes. We shall not discuss such problems in detail. However, it will enter future discussions to some degree.

6.4 True Value, Assigned True Value and Reference Laboratories.

Consider a test material which has been circulated to participant laboratories. In section 2.3 and 3.5 the assumption was made that for each specimen and each clinical parameter, there existed a 'true' value. Unless a laboratory uses an independent method of monitoring the process accuracy, it cannot check the accuracy of its own process. In fact this is one reason for participating in an inter-laboratory scheme. We would like to make the same assumption here; the test material has a 'true' value for each clinical parameter. This 'true' value is the same for each laboratory. Due to different techniques and methodologies e.g. Coulter, Technicon, etc. it is likely that this assumption will not be justified. However, the solution to the problem has been mentioned in the introduction to this chapter. We group together these participants which can be considered to be 'similar'. Similar in this situation could mean using the same device, having the same delay in receiving specimens etc. In the following discussion we shall consider a single homogeneous group. Further, we shall consider that only one clinical parameter is being investigated by the inter-laboratory scheme.

Since the 'true' value is unknown, - we have already indicated the scarcity of standards in haematology - we must agree on some value for this 'true' value. In fact we must agree on some procedure for assigning the 'true' value to trial specimens, as the problem will be encountered with each new specimen circulated. Having prescribed the 'true' value, it should be possible to make judgements concerning the reliability of laboratory work.

Two procedures for assigning the 'true' value to the circulated specimen could be considered. The value which is prescribed to the 'true' value will be called the assigned true value. These two quantities - the 'true' value and the assigned true value will not usually be equal, although we never know this, but will hopefully be close. Two procedures for setting the assigned true value are given below.

(i). The assigned true value is the mean, median or mode of the results of the participating laboratories.

(ii). The assigned true value is the mean of the results of a group of 'reference' laboratories.

Both procedures appear to be reasonable and intuitively appealing. We shall consider each in turn.

We have seen already that few reference/standards are available for calibration in peripheral blood counting, see 5.3.2. However, there is for example the cyanmethaemoglobin standard. It will be obvious that not all laboratories have the materials, equipment, manpower etc., to carry out such calibrations. Furthermore, some laboratories carry out extensive intra-quality control schemes and others do not have such extensive schemes. It would be reasonable to assume that the precision and accuracy of laboratories, even in

the same group of participants, differ. A further, non-haematological reason for such differences could lie in the method of handling and reporting results. For example, we saw in section 2.2.2 that the mean of replicate results will, on average be more precise than a single result, and we will see in section 6.9 that even in a quality control scheme where single results should be reported many participants report means.

In the following we denote by μ the 'true' but unknown value of the clinical parameter, and denote by x_a the assigned true value.

(i) Use of all participants results.

We have noted above that it is likely that for a number of reasons, the accuracy and precision of the participating laboratories will be different. Some results may be so different from the others, i.e. greatly different from x_a, that they could be assumed to be 'outliers'. We have already discussed outliers in sections 2.3.3 and 5.5. We saw that the concept of an outlier depended on the population of interest. In this situation we give the following description.

Let us assume that the results of participant laboratories have a certain specified distribution. The theoretical mean of this distribution is μ. Our best estimate of μ should be x_a. Any result which is not an observation from this distribution is an outlier. As before we are confronted with the problem that we cannot usually decide whether a particular result is an extreme value from the distribution or if it is an outlier. The basic problem with this approach is the assumption of a particular distributional form. Often as we shall see, there is an implicit assumption that the distribution is Gaussian. In many of the remaining cases there is an assumption that the results have a symmetric distribution. These

assumptions may be justified, but there are situations in which they will not be justified. Examples of both types of situations are presented by Lewis and Burgess (1969). In practice we require to eliminate, or reduce the effect of the outliers on the assigned true value. We shall discuss in turn the use of the mode, the median and the mean for the assigned true value.

The mode: The mode of a sample has been defined in section 5.5.1 (viii). Briefly the mode is the most frequently reported value. Thus, by putting x_a equal to the sample mode we are accepting the majority decision as to the 'true' value. This procedure appears to be reasonable, for we are in effect calibrating the participating laboratories. Therefore it makes sense to calibrate to the most common value, as then fewer laboratories need to modify their procedures. Two criticisms of this can be made. Firstly, there are as we have indicated, methods for determining 'true' values even if they are expensive and tedious to use. To calibrate to the majority decision, would perhaps lead to inconsistencies. Secondly, the use of the majority decision is similar to a Delphi technique. We hope by reporting the majority decision to participating laboratories, to bring participating laboratories to a unanimous decision. We note, however, there is no assurance that this will happen and further, there is no assurance that the concensus opinion will result in μ.

The median: The sample median has been defined in section 5.5.1. Briefly, the sample median is the value which bisects the ordered sample values into two equal sized parts. The use of the median as the assigned true value, x_a, is an attempt to reduce the effects of outliers. As we have noted in 5.5.1, the median is an unbiased estimator of μ only if the distribution of results is symmetric about μ. We note

that if the distribution of participants results is such that when a bias is present in calibration, it most commonly occurs in a specific direction, this would mean that laboratories without bias would almost certainly be judged 'wrong'. Figure 6.1, shows a situation in which the majority of biased results occur on the right hand side of μ. The distribution of unbiased results is symmetric about μ. But the assigned true value, the median of the sample is contaminated by the biased results.

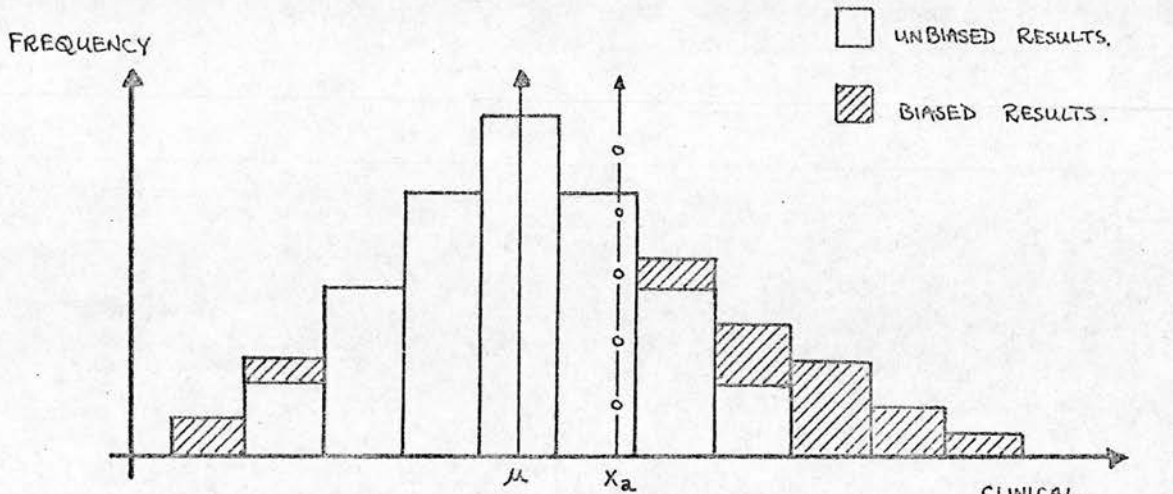


Figure 6.1.

The mean: Perhaps the most difficult to justify is the use of the arithmetic mean of the results. It suffers from the effect of outliers. Fortunately this problem has been recognised. As before, section 2.2.2 attempts have been made to censor the outliers. A common procedure, a modification of a criterion suggested by Chauvenet, (See Geigy's Scientific Tables 5th Edition), is often used. This procedure can be described in the following way.

The mean and standard deviation of all participating laboratory results, at least all within the group we are investigating, are calculated. These are sometimes referred to as 'the crude mean and standard deviation'. Any results outwith two standard deviations from this crude mean are discarded as outliers. The mean and

standard deviation of the remaining results are calculated.

Ward and Lewis (1975) found that using this procedure for results in the DHSS (LDAG) and BCSH Haematology Quality Control Trials a rejection rate of 5% is experienced. They suggest that such procedures remove 'obvious wrong results', which are usually ascribed to transcription errors. If, however, the results were so obviously wrong, we would not require this criterion to detect them. Oddly enough the rejection rate of 5% is what we would expect if the results had a Gaussian distribution and no outliers were present.

The use of the mode, median, mean or mean based on censored samples as the assigned true value results in many problems. Further, finding a solution for a particular clinical parameter and for a particular interlaboratory scheme, will not mean that the same procedure will be of use in another scheme or for another clinical parameter. Let us turn to the second proposed procedure; the use of reference laboratories.

(ii) The Use of Reference Laboratories.

By reference laboratories, we mean separate and independent institutions which implement comprehensive intra-laboratory quality control schemes, and perhaps participate in special interlaboratory schemes organised for the reference laboratories. These laboratories are judged to be highly proficient in 'the art'.

We note that these reference laboratories will almost certainly produce more homogeneous results than the group of all participating laboratories. Further, because the laboratories are 'highly proficient', the assumption is that they are more likely and more often close to the 'true' value μ.

The assumption that the results of the reference laboratories has a Gaussian distribution appears to be reasonable. Hence, we average the results of the reference laboratories. The assigned true value is set equal to this mean. (In the light of no further information, wildly discrepant results could indicate unsuitability of the circulated materials.)

The group of reference laboratories does not necessarily remain fixed. If total reliance were not to be placed on these reference laboratories, a weighted average of reference results and the results of the other participating laboratories could be used for x_a. Taken to its logical conclusion this would result in a score of reliability for each laboratory. The contribution to the assigned true value of each laboratory could be weighted by this score. One result of such a procedure is a 'league table' of laboratory capability. This has been suggested by some laboratories, see section 6.9.

Unlike the previous procedures which used all results, assuming equal quality of results, the use of reference laboratories requires the participating laboratories to meet the accepted standards of chosen 'highly proficient laboratories'. It would appear that this procedure will result in a consistent laboratory calibrating scheme. The procedure would also appear to be at least as practicable as the other schemes and present fewer difficulties.

6.5 A General Scheme for Inter-Laboratory Comparison.

We shall discuss, in this section, a general inter-laboratory quality control scheme. The discussion will show the components of variation and differences in accuracy which may arise. However, the scheme, although used in certain situations, see Mandel and Lashof

(1959) and Mandel (B27), is not practical in haematology. Having discussed the general scheme, however, will allow us to investigate modifications which may be made to give a viable design for haematology, indicating the loss in specific information caused by the modification.

A procedure and corresponding analysis for such an inter-laboratory trial has been given by Mandel and Lashof (1959) and Mandel (B27). We shall describe the basic design which is the same in both cases.

Let there be ℓ participating laboratories, and m test materials (specimens) for which each laboratory will make r replicate determinations. Only one clinical parameter shall be considered. The design can be shown schematically as in table 6.1.

Table 6.1.

Materials (Specimens) / Laboratory	1	2	3.....	m
1	xxxxxxxxx r replications	xxxxxx xxxxx		xxxx xxx
2	xxxxx xxxx	xxxxxx xxxxx		xxxx xxxx
3 . . .				
ℓ	xxxx xxxx	xxxx xxx		xxxxx xxxx

Ideally the m specimens should cover the complete range of the clinical investigation, ranging from the lowest to the highest value of the clinical parameter. In haematological situations this will simply not be possible if only for the fact that pathological blood specimens are not usually available in large quantities. The specimens are presented for measurement in a random order, although in the schematic representation the specimens are set out in ascending order of magnitude. The variability in the different laboratory results has two sources; within and between laboratory variability.

Between Laboratory Variation.

If the only variability in results was between the participating laboratories, the following analysis would allow identification of these differences. Determine the mean for each of the mxℓ cells, i.e. the mean of each of the r replications. Determine also the mean over all participating laboratories of the m specimens which have been circulated. For each laboratory, plot the mean of the r replications for each specimen against the corresponding mean for all laboratories. Ideally the graph for the i^{th} laboratory should be as in figure 6.2.

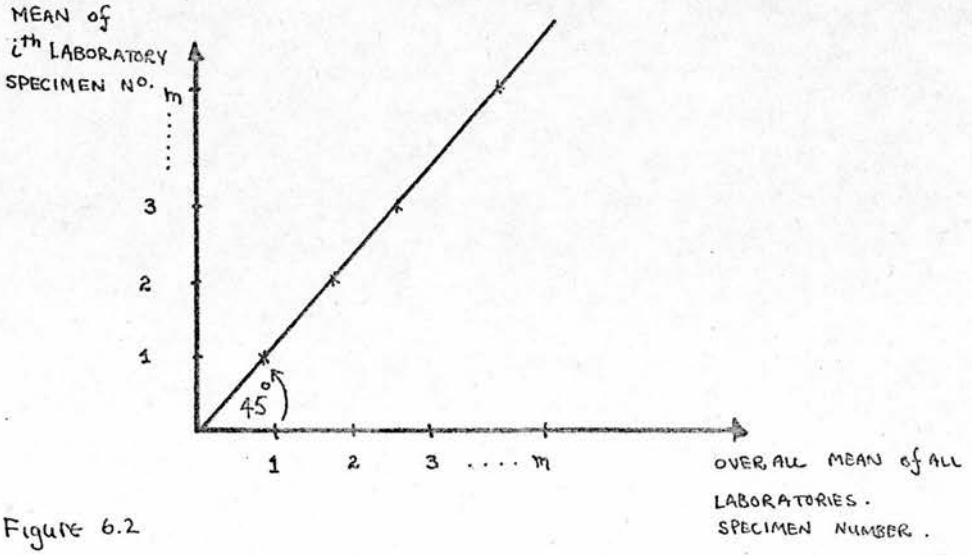


Figure 6.2

If a constant systematic difference had been present for the laboratory, the response line would be parallel to the 45^{o} line shown, but not passing through the origin. The response would be represented by one of the lines shown in figure 6.3

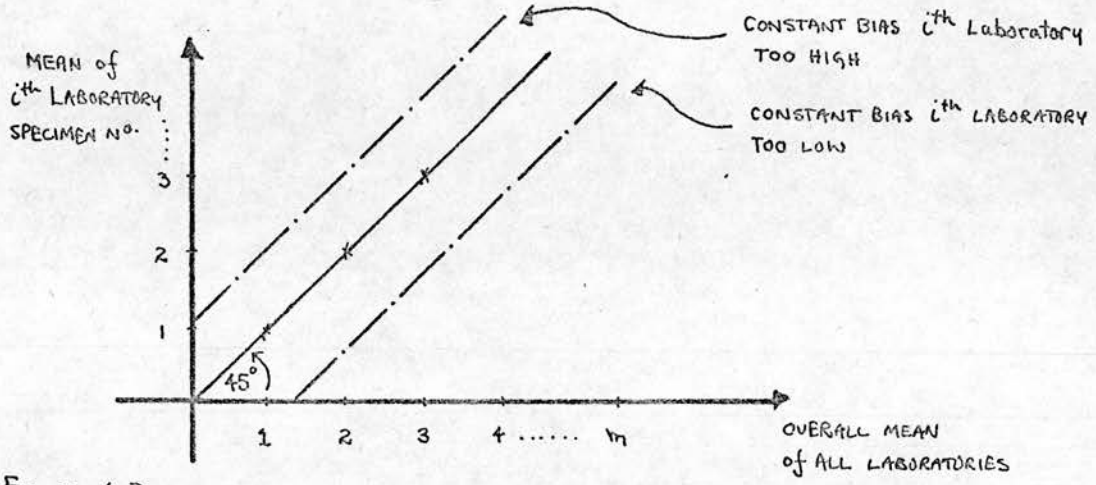


Figure 6.3

We may find, however, that the response of the i^{th} laboratory corresponds to one of the figures shown in figure 6.4.

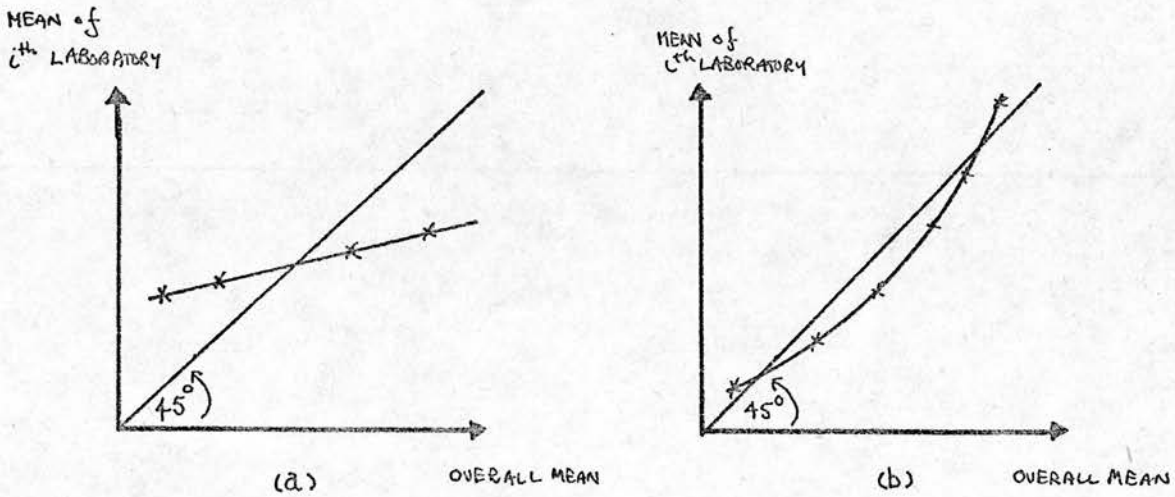


Figure 6.4

The situation shown in (a) could have been caused by both constant and proportional systematic errors while the situation shown in (b) exhibits a non-linear response and may be due to a malpractice of the measuring method.

Within Laboratory Variation.

We can identify the replication error for each laboratory. This is the precision of each participating laboratory. In practice, precision may not be constant throughout the range of clinical investigation, but we can attempt, if necessary to use a transformation to give equal cell variances.

A second within laboratory error may occur. This can be thought of as a 'scale' error. Mandel gives the following example to explain how such an error might arise.

Suppose that two samples have to be weighed. We shall call these samples A and B. The weight of A is slightly greater than one gramme and the weight of B is slightly greater than five gramme. Two standard weights are required in the subsequent weighing procedure. The true weights of A and B are known relative to the calibration of the standard weights. Thus if an error exists in the calibration of the standard weights, no amount of replication will increase the precision of such weighings. This type of error may arise chemically through interfering substances or physically through interfering properties. In the case of the Coulter-S, these physical properties could be the threshold setting or the coincidence correction automatically used by the machine.

The associated analysis of variance is carried out either on the raw data, i.e. the actual reported results, or on the transformed data, if this has been necessary. Such a design as this will allow for the identification and comparison of these different sources of error. This type of scheme, as Mandel has shown, has been used in inter-laboratory testing.

Mandel has suggested that reasonable values for m, ℓ and r, are $m \geq 5$, $\ell \geq 10$ and $r \geq 4$. The amount of data to be checked and handled quickly becomes substantial if anything other than small numbers of replications and specimens are used. We assume that the number of participating laboratories, ℓ, is large. If we consider the amount of test material which must be sent out to each laboratory, we see that a comprehensive scheme could not be implemented regularly. Certainly this is true when we consider haematology. Furthermore, for some laboratories, the amount of work involved could not be justified. The cost in terms of time, administration by both laboratory and organising staff would, in all probability, be prohibitive.

Although we have seen that the scheme above would allow for identification of the cources of variability, in practice, as we have seen, the scheme could not be implemented. Let us consider how we might reduce the size of the scheme. Three reduced schemes will be discussed. However, by reducing the size of the scheme we lose information and so we shall indicate the loss of information in each of the modifications.

(a). A single test specimen is sent to each of the participating laboratories, i.e. $m = 1$. Each laboratory measures the specimen once only, i.e. $r = 1$. If all laboratories report a 'similar' value, then we can be relatively satisfied that no major differences exist between laboratories at the particular level of the clinical parameter exhibited by the test specimen. This does not mean that there are no differences between laboratories. If discrepancies in the reported results exist - the measurement of discrepancy relies on more than the information contained in one such trial, we have no information as to what causes the discrepancies. Hence, no help can be given to the participating laboratories.

(b). We modify (a), requiring each laboratory to make a number of replicate measurements on a single test specimen, i.e. $m = 1$ and $r > 1$. In this case, we can obtain an estimate of the precision of each laboratory. We should be wary of such estimates of precision. In section 2.3 we saw that certain laboratory rules-of-thumb for rejecting replicate measurements affect the estimate of variability of the process, i.e. the laboratory's precision. The use of this procedure would lead to a failure to detect and identify within and between laboratory variability.

(c) Let us suppose that two test specimens are circulated to each laboratory, i.e. $m = 2$. Only one measurement is made on each of the specimens, $r = 1$. This design makes available information concerning individual laboratory biases. The information could be analysed by the usual two-way analysis of variance (making the assumption of equal laboratory precisions). In fact as we shall see this design is used. A graphical method of displaying the results has been described by Youden (1959) and modified by Tonks (1963) see 6.6. By using this design, we usually require to assume constant laboratory differences, and also equal laboratory precisions at each specimen level.

6.6 The Youden Plot and Modifications.

This scheme has been extensively used in inter-laboratory trials. We give a brief description.

Two specimens A and B are sent to each participating laboratory. For each clinical parameter of interest, one measurement is made. (For a device such as the Coulter-S, this means that the specimen is

passed through once, and the necessary parameter values are taken from the single report form.) Youden (1959), in a description, has specified that A and B should be reasonably close to each other in the magnitude of the clinical parameter to be measured. This is required to avoid complications which could arise from 'differential behaviour of the test materials'. Elion-Gerritzen (1977) has suggested that it is now customary to use two specimens of different magnitudes although 'similarity in composition has, understandably, remained unchanged'. Elion-Gerritzen specifically considers situations arising in clinical chemistry. A further suggestion is that different concentrations could be used.

We note that in haematology, interest lies often in pathological blood specimens. It is unlikely that enough 'abnormal' blood could be obtained to circulate participant laboratories. There is the possibility that animal blood could be used to simulate certain pathologies. However, we must operate under the assumption that there is no differential behaviour of the test materials.

The presentation and analysis of the test results is an important feature of Youden's procedure. The result of the i^{th} laboratory, $i = 1,2,\ldots,\ell$ is denoted (a_i, b_i). The measurement made on the specimen A is, a_i, while b_i is the measurement made on the specimen B. These results are plotted on a graph which has the same vertical and horizontal scales, which are often in terms of standard deviations (either of the collected (censored) data, or of reference laboratories) see later. The median lines are drawn. The median line vertically, is the line which gives the same number of points to the left as to the right, and is parallel to the vertical axis. The median line horizontally, is the line which gives the same number of points above as below, and is parallel to the horizontal axis. Such a graph is shown in

figure 6.5.

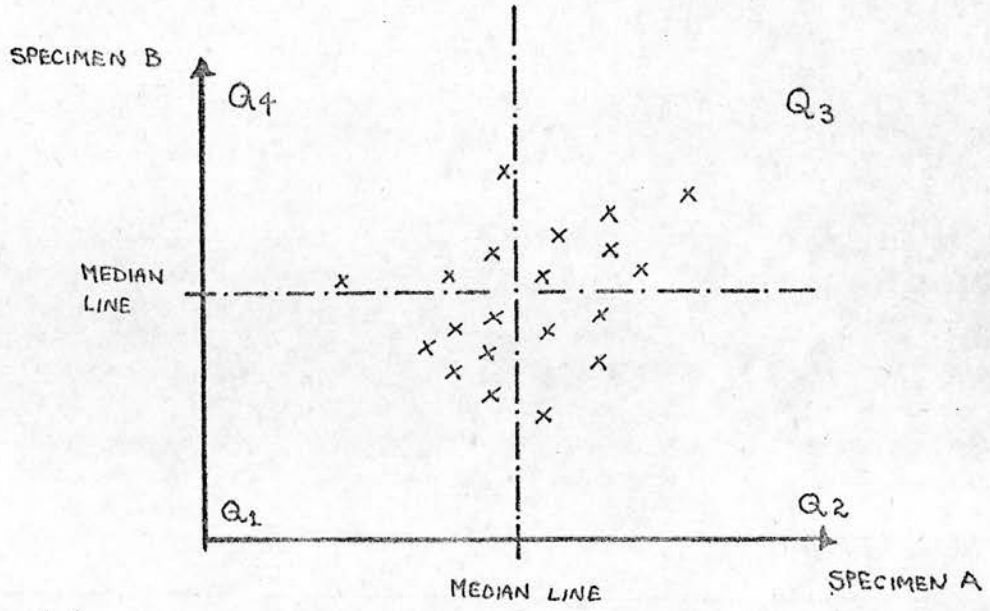


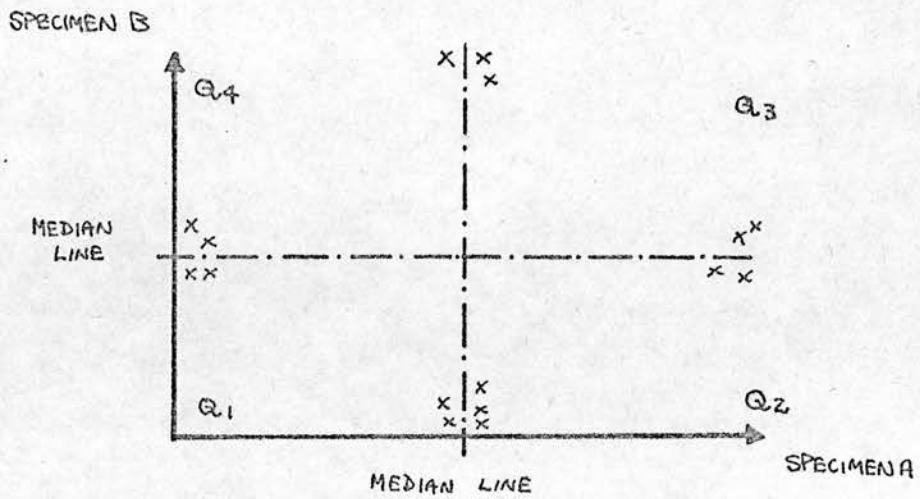
Figure 6.5

We note that the proposal that reference laboratories are used implies the replacing of the median lines by the corresponding lines representing the means of the reference laboratories.

The scatter diagram, figure 6.5, is divided into four quadrants Q_1, Q_2, Q_3 and Q_4 by the median lines. Certain configurations of points can indicate the 'abilities' of the participating laboratories. A random scatter through all quadrants, but centred on the intersection of the median lines is referred to as the 'ideal' situation, being most likely caused by random error. If the scatter is widespread, it could be due to heterogeneity in sampling methods.

A concentration of points in quadrants Q_1 and Q_3, close to the 45^o line through the intersection of the median lines, a common situation, indicates a tendency of the laboratories to have low results in both specimens or high results in both specimens. There is a consistency in such laboratories. If the concentration of points is in quadrants Q_2 and Q_4, this could mean inconsistency in the laboratories. If points occur as shown in figure 6.6, it is likely

that transcription errors have been the cause.



If the specimens have similar clinical parameter magnitudes, thus allowing for the assumption of equal precisions for both specimens we can by assuming also that the participating laboratories have the same precisions, obtain an estimate of the precision.

An estimate of overall laboratory precision can be obtained by considering the perpendicular distance of each point from the 45^{o} line through the intersection of the median lines. Alternatively, let a_i and b_i be the measurements on A and B made by the i^{th} laboratory and let $d_i = a_i - b_i$, $i = 1,2,\ldots,\ell$ and let $\bar{d} = \frac{1}{\ell} \Sigma\, d_i$, then an estimate of laboratory precision is given by s or s'.

$$s = \frac{\sqrt{\pi}}{2} \frac{\ell}{\Sigma} \left|d_i - \bar{d}\right| \quad , \qquad s' = \sqrt{\frac{\Sigma d_i^2 - \ell.\bar{d}^2}{2(\ell-1)}}$$

Having estimated the precision of laboratories, a circle of radius 3s is drawn with centre at the intersection of the median lines. If the graph has been scaled in standard deviation units, the radius is simply 3. Laboratories whose reported values lie within this circle are regarded as operating acceptably. The other laboratories are judged to be unacceptable.

A number of authors haye suggested that the circular acceptance region, described above, should be modified. Basically the modified regions have been elongated in the axis of the 45^{o} line through the median point, e.g. see Bailly (1975), Van Assendelft and Holtz (1975). Such modifications are made in consideration of the possibility that the pairs of measurements from the participating laboratories are not independent.

Elion-Gerritzen has commented on Youden's use of the term systematic error, pointing out that systematic error can be consistent, proportional or non-linear. On a Youden plot, if the majority of the laboratories show constant differences, a proportional effect would be deemed to be a random error. The suggestion is to plot the graph in a real scale and not in a coded standard deviation form.

The major advantage of Youden plots, is that they are readily understood by all or nearly all members of the laboratory staff. However, we remark that the plots should be used to give an indication of a laboratory's performance and not as the basis of significance tests. Such inferential procedures, e.g. acceptance regions etc., are probably not particularly valuable as they depend on assumptions which may not be totally justified due to the multitude of factors which can affect the trial results.

The distributional problems met in the above, e.g. lack of independence, can be circumvented by appealing to non-parametric methods. A number of ranking methods have been reported, but we shall not discuss their use here.

6.7 Scheme of Zender and Linder.

A criticism of the Youden treatment is that it really only monitors the accuracies, at two points, of the participating laboratories. The estimates of precision which have been given are at best only a rough indication of the overall (or particular) laboratory precision. This is not a problem of analysis, it lies in the design. We have seen that the general scheme discussed in section 6.5, cannot in practice be used but in reducing the size of the scheme in this way (i.e. 2 test specimens & 1 replication each), we sacrifice information on laboratory precision.

Some attempts have been made to obtain information concerning precision from such a two sample, single replication scheme. The problem is that in trying to obtain, and in presenting the information, the most attractive property of inter-laboratory schemes is lost; namely the simplicity. Further, the inferences from such analyses are dubious. We shall discuss briefly one attempt by Zender and Linder (1975).

The basic situation is similar to that described for Youden Plots. Two specimens A and B are circulated to each of the participating laboratories, ℓ in total. The results reported by the i^{th} laboratory are a_i and b_i. Let the true, but unknown values of specimens A and B be α and β respectively. As before, we require to use assigned true values. In haematological situations these will most often be group means (censored) as described previously. The following transformations are used on the original data.

$$c_i = \frac{\alpha}{\beta} b_i, \quad z_i = \tfrac{1}{2}(a_i + c_i), \quad d_i = a_i - c_i \qquad i = 1,2,\ldots,\ell.$$

Zender and Linder state that if precision of the i^{th} laboratory is excellent, then c_i and a_i will be approximately equal. If accuracy is good, then d_i will be approximately zero. Two test statistics are proposed for testing laboratory accuracy and precision. The basis of these tests are the sums of squares S^2_{zi} and S^2_{di}.

$$S^2_{zi} = \frac{1}{\ell-2}\left(\sum_{i\neq j} z_i^2 - \frac{\left(\sum_{i\neq j} z_i\right)^2}{\ell-1}\right)$$

$$S^2_{di} = \frac{1}{\ell-2}\left(\sum_{i\neq j} d_i^2 - \frac{\left(\sum_{i\neq j} d_i\right)^2}{\ell-1}\right)$$

To test the i^{th} laboratory's accuracy, it is proposed that we use the statistic

$$F_a = \frac{(z_i-\alpha)^2}{S^2_{zi}}$$

To test the i^{th} laboratory's precision, it is proposed that we use the statistic

$$F_p = \frac{d_i^2}{S^2_{di}}$$

Under the assumption that there are no differences between laboratories (constant and/or proportional, calibration differences, and differences in precisions) the F_a and F_p have F-distributions with (1, ℓ-2) degrees of freedom. (The truth of these distributional statements depends on a number of assumptions being justified e.g. that measurement errors have Gaussian Distributions, that the precision of laboratories are the same at the different clinical levels α and β etc.).

Now let us consider the suggested analysis. Firstly even if all the distributional statements were correct, testing at the 5% level in 2ℓ tests (ℓ for precision and ℓ for accuracy) we can expect $\ell/10$ tests

to produce significant results even when no differences exist. Secondly, if $\alpha/\beta \neq 1$, the variables d_i and z_i have covariance $\frac{1}{2}(1 - \frac{d^2}{\beta^2})\sigma^2$. This means that the results of the tests based on F_a and F_p are dependent. Thus we do not usually have independent assessments of a laboratory's accuracy and precision.

Basically, the presentation and analysis suggested give no better understanding of the situation than Youden Plots. A better procedure would be to simply use the Youden Plots and accumulate information on the past performances of the specific laboratories. The cumulative information can be used to obtain an overall impression of the laboratory's performance. For example the 'average distance' from the target values of a specific laboratory gives a measure of that laboratory's total error. Further, any persistent bias can be detected.

6.8 The DHSS (LDAG) and BCSH Haematology Quality Control Trials.

A comprehensive description of these trials is given by Ward and Lewis (1975). A paper by Lewis and Burgess (1969) describes the initial trial and suggests how they might develop in the future.

Two specimens, X and Y are sent to the participating laboratories, usually at three to four week intervals. The laboratories are asked to report their peripheral blood counts on both specimens. Occasionally, special investigations are circulated, e.g. serum iron, abnormal haemoglobins. The results from each laboratory are grouped according to the method of determination - automatic, semi-automatic or manual, and also according to the time delay in receiving the specimens for analysis.

The stated objectiyes of the trials are:

(i). To investigate materials for use in standardising haematological measuring processes.

(ii). To show overall precision of methods based on the results from fully automated laboratories, which have analysed specimens unaffected by time delays.

(iii). To provide each participating laboratory with an index of competence.

For each group of results, and each clinical parameter, a crude mean and standard deviation is calculated. All results outwith two standard deviations from this mean are rejected. The mean and standard deviation of the remaining results are calculated. (See Chauvenet's criterion section 6.4) We denote these by $\bar{x}$ and s.

For each specimen, X and Y, and for each clinical parameter, a variance index (ratio) is determined. The variance index, for a specific clinical parameter, for a measurement x_i made by the i^{th} laboratory, is denoted $|R_x|$ and is

$$|R_x| = \left|\frac{x_i - \bar{x}}{s}\right|$$

When both specimens results have been accepted, i.e. not rejected by Chauvenet's criterion, for each parameter the ratio $z = x/y$, $(x < y)$ is computed. In a similar way, a variance index $|R_z|$ is computed for these ratios.

A further two variance indices are computed. Firstly, for each laboratory the index $|R_i|$, the investigation index is determined for each clinical parameter:

$$|R_i| = \frac{|R_x| + |R_y| + |R_z|}{3}$$

The second index is the overall variance index for a specific laboratory, $|R_o|$. This is simply the mean of all the indices $|R_i|$.

The same scale of assessment is used for all variance indices. The scale is given in table 6.2.

Variance Index	Assessment
less than 0.5	Excellent
0.5 - 1.0	Good
1.0 - 2.0	Acceptable
greater than 2.0	Unacceptable

Table 6.2

Let us look more closely at the characteristics of the 'variance index' which features so prominantly in the trial assessments and presentation. Suppose we consider a single clinical parameter X and obtain the determinations, $x_1, x_2 \ldots, x_n$, from n participating laboratories on a single trial specimen. Let us assume that all outlier detection and deleting procedures have taken place. The variance index for the i^{th} laboratory is,

$$|R_x| = \left|\frac{x_i - \bar{x}}{s}\right|$$

where $\bar{x}$ and s are the mean and standard deviation as defined in section 6.4. Firstly, we note that the use of the modulus destroys information on the actual 'bias' of the i^{th} laboratory from the target value $\bar{x}$. This is important information especially over a series of trials. A consistent sign would represent a consistent 'bias'. A variable sign, with a large scatter in the numerical value of $|R_x|$ might indicate a serious defect in precision. It is appreciated that this information

can still be obtained from the results presented. However, the participant laboratory must carry out the required computations.

Let us assume that the data x_i, $i = 1,2,\ldots,n$ have a Gaussian distribution. Then the percentage of laboratories, on average, with specified variance scores can be computed, see table 6.3.

Table 6.3

Variance Index	Expected Percentage of Results
0.0 - 0.5	38.29
0.5 - 1.0	29.98
1.0 - 2.0	27.18
greater than 2.0	4.55

From table 6.3 we can calculate the approximate expected value of the variance index. This is 0.83, (theoretically the expected value is $\sqrt{\frac{2}{\pi}}$.)

We cannot monitor 'increase in quality' using this index. $|R_x|$ is a standardised score, i.e. it is based on the mean and standard deviation of the particular data in the particular trial. Thus, if the basic shape of the distribution of the participants' results remains relatively stable, the same average score is expected, whatever change may take place in precision. This point is supported by two pieces of information. Firstly, from a report circulated to participants in 1973 (January), it was seen that the average variance indices all appeared to be slightly less than 1 over a series of trials. The second indication is given in table 21, section 6.9. The mean variance index for 'fully-automated', 'semi-automated' and 'Manual' are 0.87, 0.88 and 0.85.

If our proposals, section 6.4, were implemented, i.e. reference laboratories are used to establish the target values for trial specimens, the denominator of the variance indices would be the precision of the reference laboratories. This is much more likely to be constant, excepting for unsuitable test materials. Further, it assesses laboratory's performances using an attainable and accepted precision.

We note also that the overall index and the investigation index can be based on different numbers of indices. This leads to a problem with comparison between laboratories, using such indices. Further, the simple average index, $|R_o|$ implies that all determinations reported have the same clinical importance to all participating laboratories, which in practice is probably not the case.

Cumulative information is also kept. The cumulative indices $|R_c|$ and $|R_s|$ store information relating to the mean and standard deviation for each method for each participant. Biennial reviews are given.

A visual representation of a laboratory performance compared with all participants results can be given. Unfortunately, in practice this has not been done. The procedure (graph) has been called the 'Ratio Test'. Basically the graph is the same as that shown in Figure 6.2, i.e. a specific laboratory's results are plotted against the National, (overall) means. If the two plotted points lie on the 45^{o} line through the origin, we can conclude that the laboratory is operating correctly, at least for these specimens. If the two points lie on a line through the origin which does not have 'slope' 1, (the same scales are used on both horizontal and vertical axes), we suspect a consistent standardising error. If the graph shows some other type of situation,

we cannot say with certainty what causes the 'unreliability' of the specific laboratory.

We note that the basic set-up of the trial would admit a presentation such as Youden Plots. (In fact a number of respondents requested such a visual display, see section 6.9 table 25(b)). Perhaps it would be best if a Youden plot was presented in the trial report and further each laboratory was directed to sketch the 'Ratio Test' graph described above. (This graph is easily constructed.) By using these graphs, the specific laboratory's results and the National average results, a reasonably detailed picture of the laboratory's performance can be gained. The information so gained will probably be as meaningful if not more so than that gained from the somewhat obscure and complex 'variance indices'.

As we have stated, the basic scheme in these trials is similar to that discussed when considering Youden plots. However, the method of 'analysis' is, at least to the average laboratory, much more complex. In fact from what we have said concerning the characteristics of the variance index, the presentation is probably less meaningful to participants than a reasonably straightforward graphical presentation. We shall comment further on this point when we present information obtained from a survey of participants. A brief description of this survey follows with the results presented in tabular form.

6.9 Survey of Participants of the DHSS (LDAG) and BCSH Haematology Quality Control Trial.

In section 6.8, we discussed the general form of the DHSS (LDAG) and BSCH Haematology Quality Control Trial. The aims of the trial were also given. In the discussion of inter-laboratory trials which

we have presented, we have stressed the importance of each participant clearly understanding what is expected of the commitment to the trial.

This requirement, that is the participation, has two facets. Firstly each participant should be aware of the aims of the trial in both the short and the long term. Further, the participant should also be aware of its own short and long term aims in participation, (if any). Secondly, the participant must be aware of the trial procedure, both administration and haematological methodology, and the necessity for strict adherence to it.

Basically, the reviews of various interlaboratory trials which have appeared in the literature have been by investigators either with vested interests or by people with higher than average understanding of them. Much of the investigation has centred on the 'success' of the trials in terms of what they have shown concerning the overall laboratory competence.

The information which was desired from this survey concerned the participants of the trial. The survey was designed to ascertain the extent to which participants understood the aims of the trial, followed the procedure laid down, used the reported trial results and what changes, if any, they felt should be made. (It was appreciated that unless a 100% response rate was achieved, the information collected could be biased, e.g. perhaps only the more interested laboratories respond etc.)

A preliminary draft questionnaire was constructed and subsequently vetted by my supervisors, Dr. R.J. Prescott and Dr. S.H. Davies. Minor changes were made. These changes mostly related to vocabulary - an effort to decrease the chance of a respondent misunderstanding a question.

The draft questionnaire was sent to Dr. S.M. Lewis and Mr. P.G. Ward of the Royal Postgraduate Medical School, London, who had had previously been contacted to enquire if they were agreeable to such a survey being carried out. The draft was promptly returned with a request for the addition of questions 23, 24 and 25. Some minor comments were made, again regarding the vocabulary which had been used.

At a meeting in Edinburgh with Dr. Prescott, Dr. Davies, Mr. Newman (Chief Technician, Haematology Department Royal Infirmary Edinburgh) and myself, the final form of the questionnaire was established, see appendix 3.

Attached to the questionnaire was a covering letter, see appendix 4. The questionnaire and covering letter were reproduced and sent to the Royal Postgraduate School, London. The distribution of the survey was effected through the normal channels for distribution of trial specimens and accompanied the test materials for October 1977.

The questionnaires were returned on completion to London and thence forwarded to Edinburgh. (A closing date of one month was set.) This method of distribution and return, as explained in the covering letter, see appendix 4, was to enable a guarantee of anonymity to be given to the participating laboratories. (That is to protect the DHSS trial codes.)

A letter from the DHSS was forwarded, by the next trial, to all participating laboratories. The letter is reproduced in appendix 5.

The questionnaire had two sections, as described in the preamble to the questionnaire, see appendix 3. The first section contained questions relating to actual laboratory practice. It was felt that such questions, important and informative in their own right, allowed a clearer impression of laboratory practice to be obtained than relying on information gleaned from responses in section B. Section B

contained questions relating to the administration and procedures within the laboratory necessary for participation in the DHSS (LDAG) and BCSH Haematology Quality Control Trial. The final questions in this section prompted responses to behavioural questions, and also invited comments pertaining to the Trial in general.

We present the responses to the questionnaire in tabular form at the end of this chapter. A discussion and commentary on the more interesting aspects of the responses is also given.

Of 420 laboratories circulated, two laboratories were circulated with two questionnaires while a third received four. These laboratories receive more than one Trial report, i.e. they receive more than one Trial package of specimens. As we indicated the survey was distributed from London using a 'report circulation list'. For this reason, a further nine laboratories which responded only receive the Trial reports and had been asked to respond to questions 22, 23, 24, 25, 26, 28 and 29 and the general comment only. Further, two laboratories did not respond indicating that they had now withdrawn from the Trial. In all 300 laboratories returned completed, (or partially completed) questionnaires. This represents a response rate in excess of 70%.

In the following discussion the table numbers correlate to the question numbers of the questionnaire (see Appendix 3).

The number of peripheral blood counts made, on average, per day was thought to give a reasonable indication of the 'size' of the responding laboratory. The frequency distribution presented in table 1 shows that the respondents came from a cross-section of sizes. The number of laboratories classified as '50 or fewer', contained a number of research laboratories, some performing determinations on animal blood. From the general profile, it would appear that no apparent bias is present with respect to the size of responding laboratories.

Since we have discussed the concept and problems connected with 'normal ranges' and since they seem to be used in practice, an indication of the source of the 'normal ranges' used in each laboratory was invited. This is, as far as we can ascertain, the first survey of the use of 'normal ranges' in haematology. Table 2 shows that apart from MCV, 60.5% of respondent employ the 'normal ranges' of Dacie. However, it is noticeable that laboratories do not necessarily use the same source 'normal ranges' for all clinical parameters. This is particularly noticeable in the case of MCV. We note that few, about 8% of laboratories, employ the Manufacturer's 'normal ranges'.

Table 3 shows the response to the question 'Have you an established formal laboratory quality control scheme for blood parameters?'. By 'formal', we had meant a written procedure, explicitly outlining quality control procedures and checks which are made in routine practice and further giving corrective procedures for out of control situations. It is not clear that this was the meaning which the respondents took, especially from comments made on the questionnaire beside this and other questions. Few laboratories, about 6½%, claimed to have no formal quality control scheme.

Over 90% of responding laboratories monitor the accuracy of their measuring process. It is not clear what the other laboratories do. Presumably they must calibrate at machine set-up and have faith that this is maintained throughout. From comments in the questionnaire margin, it is apparent that some laboratories do not use accuracy and precision in the way defined in sections 3.5.1 and 3.5.2.

The materials used in the monitoring of accuracy, and the number of laboratories employing them are shown in table 5(a) and 5(b). Firstly, we note that laboratories tend to use more than one

material to monitor the accuracy of their measuring process. The majority, over 90%, of laboratories use a commercial material. We shall comment on this point at a later stage. However, we note that since, as we show in table 16 about $\frac{2}{3}$ of the respondents employ a Coulter-S, this suggests that a majority of laboratories calibrate using 4C. Even if other laboratories are not compared directly with the 'automatic' results of these laboratories, because of a wish for consistent laboratory results from a single establishment, the Coulter-S users will tend to make comparisons with their other method of measurements. Hence indirectly 'manual' and 'semi-automatic' reporters are compared against the 4C users. Because of the problems we have discussed in section 6.4 the assigned true value in any method of determination will be influenced to a greater or lesser extent by the reliability of the 4C preparation. (Comments relating to some of these points, have been made by Prangnell and Johnston (1976) and S.M. Lewis (1976)).

It is only possible to use the DHSS trial results as a retrospective monitor of accuracy due to the time delay between analyses and receiving the results. Hence the large number of laboratories claiming to use the Trial results to monitor accuracy must surely mean in retrospective assessments.

We note from table 6 that the percentage of laboratories monitoring precision is over 90%, which is smaller than that monitoring accuracy. As expected no laboratory reported monitoring precision without also claiming to monitor accuracy.

The materials used in the monitoring of precision, and the number of laboratories employing them are shown in table 7(a) and 7(b). Again we note the high frequency of laboratories employing a commercial preparation. As expected there is an increase in the use of donor

blood compared to monitoring accuracy.

Question 8, which referred to participation in a commercial laboratory quality control scheme was included, since it was felt that such schemes may in the future expand, especially for the small or medium sized laboratory. We note that 12% of participants report taking part in a commercial scheme.

From table 9(a), we see that approximately 13% of laboratories admit not using any statistical techniques. Of the remaining laboratories, the majority at least carryout a 'visual check' on the process table 9(b). It is interesting following our discussion of section 5.4, to note that approximately the same number of laboratories employ the Shewhart chart as those using the Cusum Chart. It is comforting to see that only 5 laboratories employ Pragnell and Johnston's modal control method, see section 5.6.4.

It is worth noting that only 56% of laboratories report using calculating machines in their quality control. By calculating machine, we mean non-computer type devices. It is also worth noting that only a very few laboratories claim to use computer facilities.

Table 11 shows the frequency of the monitoring of carryover and linearity. Firstly, it appears that about 26% of responding laboratories regard linearity as 'not applicable'. This is a somewhat puzzling response. Presumably linearity applies to all measuring processes. (We have noted in chapter 3, that linearity is difficult to monitor and perhaps this accounts for this response.) The table shows that most of these laboratories are in the class - smaller than 100 peripheral blood counts on average per day. Secondly we note that about $\frac{1}{3}$ of the laboratories only check linearity and carryover on the manufacturer's service. If as is usual, this is about every three to four months, we

should point out that from our discussion concerning carryoyer, see section 3.5.3, this would appear to be too infrequent. To put our comment regarding carryover in perspective we see that from table 12 that about 70% of laboratories report in the automatic device category.

Ward and Lewis (1975) reported that 350 laboratories were participating in the Trial at that stage. As we have noted 420 laboratories were circulated with this questionnaire. Table 13 shows that 88% of responding laboratories report participating in the Trial for more than two years. The correspondence in these figures appears consistent enough to suggest that we have no appreciable bias in the number of 'new' and 'old' laboratories to the Trials.

An inconsistency appears to arise in the fact that about 59% of respondents report taking part in all trials, see table 15. Ward and Lewis (1975) have stated that:

'Most laboratories fail to report one or two trials in a year'. However, our findings are supported from information reported in 1973 (circulation to participating laboratories). In this report, Lewis and Ward reported that 160 out of 264 laboratories returned all trials received.

From table 16 we note that 65% of laboratories use a Coulter-S. However of laboratories reporting in the fully automatic class there are 95% using a Coulter-S. A respondent pointed out that it may be 'fairer' on Technicon users to group them in a separate class. We must expand the point made in accepting it. We have indicated that a concensus value for trial specimens should be sought. In other words an aim of these trials should be that eventually all laboratories have the same calibration. If this aim is accepted as paramount, all participants should be judged together e.g. Coulter, Technicon and manual. However, if we are to indicate to participants whether

accuracy and/or precision are deficient, we must estimate the precisions within the different methods and methodologies. In such a way we can separate the total variability of responses into machine/method precisions and bias between machines/methods.

We have already discussed the effects of averaging replicate determinations, see section 2.2.2. In short the precision of an average of replicate measurements is on average higher than a single determination. We note from table 17, that only 57 out of 290 responding laboratories did not carry out replicate determinations. Thus of these laboratories 233 broke the prescribed Trial procedure. However, we note that although 233 carry out replicate determinations, of these 147 actually report an average. The 86 laboratories who do not report an average must however also be thought of as violating the Trial procedure since in practice, wildly discrepant results may well result in non-participation in that trial. We conclude that any inference concerning the precision of participating laboratories must be considered to indicate a situation better than reality. Hence the Trials do not reflect routine laboratory precision. Emphasising this point is the indication that 15% of laboratories make the Trial determination after calibration and resetting, while 6% make the determination after a full cleaning. This will also tend to give a picture which is better in terms of both accuracy and precision than can be attributed to routine specimens, see table 19. To complete the indication that the Trial measures a better than routine situation we see from table 20 that about 64% of laboratories treat the Trial specimens as special specimens, although they are handled by the day's operator. A number of respondents point out that this is partly due to the fact that the Trial specimens arrive in 'different containers' from the routine specimens. Thus to make sure that specimens are treated as routine

specimens the laboratory would have to decant into a 'routine container'. This particular issue cannot be settled by the Trial organisers, as the administrative problems leading to a solution would be insurmountable.

From the table 21, we see that by far the majority of laboratories report that the average of their past five results is classed as 'good' by the Trial variance index. As we have already pointed out, (even allowing for possible bias), this result is to be expected due to the nature of the variance index, see section 6.8.

We note from Lewis and Burgess (1969) that a number of original aims were set for the Trials. Of course it was indicated that the Trial would not necessarily retain all of these aims as it evolved. Among the main aims were to investigate methods for carrying out inter-laboratory quality control schemes, to investigate different controlling materials and the monitoring of laboratory accuracy and precision. Unfortunately Ward and Lewis (1975) do not include in their description of the Trials a list of aims. Although Lewis (1975) and together with Burgess has emphasised the need for investigation of control materials, only 76% of laboratories thought that this was an aim of the Trials. Further, 34% of laboratories thought that the monitoring of specific laboratory capabilities was an aim. In fact, although the Trial results report the particular results of a laboratory this is to enable the laboratory to identify problems with control, if any.

Ward and Lewis (1975) have written:

'This scheme has become well established, and participants have become familiar with print-outs.'

However, in the very next sentence the following statement is made.

'Frequent program revision has been necessary, particularly for program 08, which prints the trial reports, so that the information sent to participants can be made as clear and as concise as possible.'

It would appear that these statements are somewhat contradictory. Further, from table 23 we see that in all 72% of laboratories experience some degree of difficulty in understanding the trial presentation, I think that we should not look for 'familiarity' with print-outs, but understanding. Thus, it appears that a fundamental aim of the inter-laboratory trial is not fulfilled, namely that participants should understand what the results mean.

The most disturbing feature of this, is that 83% of laboratories have never requested help in interpreting results. Further we note that of the laboratories seeking help only 50% requested help directly from the DHSS unit. Many respondents in this category, commented that they had found the DHSS staff only too obliging.

From table 25(a) we note that 54% of laboratories, indicated that they wished no change of the results format. There is a rather bewildering feature if 83% of laboratories experience some form of difficulty in interpreting them in their present form. Perhaps, participating laboratories believe that results from such inter-laboratory trial are difficult to interpret no matter what the result format is. Of the laboratories indicating that a change would be desired, we see from table 25(b) that the majority wanted a simpler, perhaps graphical presentation.

The majority of laboratories responding appear to be satisfied with the frequency of the trials. (75%) table 28. It is of interest to note that 22% of the laboratories indicated that the frequency should be increased. There was an indication from some respondents

that the frequency should be increased when the scheme now in operation permits a quicker return of laboratory results.

When asked if the trials should be expanded in content, 56% of laboratories indicated that they should table 29(a). Table 29(b) presents the clinical parameters which were suggested for inclusion. Of the laboratories wishing for more clinical parameters, by far the majority asked for a platelet count to be included. The next most popular requests were for serum vitamin B_{12} and serum folate determinations. It was noticeable that certain laboratories obviously wished to include clinical parameters in which they were particularly interested, but which are not necessarily measured in all the laboratories taking part in this particular scheme.

Forty six per cent of laboratories claimed to use Trial results in their quality control table 30. Many other respondents stated that they would use the results if only they could receive them sooner. To this end local haematology quality control schemes were better. We note that this is only to be expected.

From table 31(a) we note that 25% of laboratories take part in other peripheral blood count inter-laboratory schemes, while 52% of laboratories took part in some other inter-laboratory trial table 31(b). The participation in other schemes should perhaps lead to more confidence and better understanding in the participating laboratory. However, it may be that laboratories require to be educated in inter-laboratory trial procedures.

It is of interest to note that 55% of laboratories indicated that they had never experienced a series of 'bad trial results' table 34. Of the remainder many stated that the trial information was practically worthless, due to the delay in receiving the information and that their

own quality control schemes had indicated problems with the measuring process very much sooner.

The overwhelming comment made concerned the retrospective nature of the results. However some other interesting points were made. Many were concerned by the possibility of producing a 'league table' of participants. However, probably just as many welcomed the anonymity and lack of an 'expert panel'.

A number of respondents hoped for better documentation for the trials. Others specified changes in the reported results e.g. the inclusion of C.V. etc. The other type of response was essentially a request for information e.g. is there a container effect on MCV?, do laboratories with computer facilities do consistently better? etc.

We conclude by saying that we believe inter-laboratory quality control schemes are of importance. There are many reasons for their importance in haematology. Not least of these is the problem, already discussed, concerning the lack of suitable standardising materials. The scheme initiated by Ward and Lewis which has evolved in to the DHSS (LDAG) and BCSH Haematology Quality Control Trial has made a significant improvement in this field. However, the following points should be made. There appears to be 'confusion', lack of adherence and misunderstanding concerning the trial procedure e.g. reporting of averages, special specimens etc. This may well arise from a desire to perform well in the trial. More likely, it arises from a lack of information (that does not necessarily mean that information has not been sent). Further, if the aims of the trial are to reflect current routine quality, this quite clearly has not been achieved.

The second major point to be made concerns the feed back of information to the participating laboratories. We have seen that a

majority of laboratories experience some difficulty in understanding the results of the Trial. From our discussion concerning the variance index, we would suggest that this measure, at least in its present form does not give more information rather than a simple graph e.g. Youden Plot. (The situation is changed if reference laboratories are used.).

We finish by noting the large response to our survey (at least 70%), and also the many varied 'extra' comments made by respondents. This would clearly indicate the desire for a good, understandable, quick national inter-laboratory trial. It is to be hoped that the DHSS (LDAG) and BCSH trial has laid the foundation for such a scheme.

TABLES OF RESPONSES TO QUESTIONNAIRE DHSS BLOOD TRIALS

Section A : LABORATORY PRACTICE

1. On average, how many routine blood count specimens does your laboratory handle on a normal weekday?

No. Peripheral Counts	No laboratories
fewer than/or 50	48
51-100	38
101-150	40
151-200	49
201-250	47
251-300	31
301-400	18
401-700	10
more than 700	10
	291

2. How are your 'Normal' blood ranges derived?

	RBC	WBC	Hb	MCV	MCHC	PCV/Hct
Dacie (Modification)	8	8	8	10	9	8
Own	50	52	54	67	59	56
Manufacturer	21	19	21	27	21	25
Dacie	187	189	186	159	176	178
Wintrobe	6	5	6	5	5	5
Own (Modified Manufacturer)	5	5	5	6	5	5
Dacie (Modified Manufacturer)	3	3	3	3	3	3
Richard/Williams	2	2	2	2	2	2
Dacie and ICSH	1	1	1	1	1	1
Eastham	1	1	1	1	1	1
Dacie/Wintrobe	1	1	1	1	1	1
Paediatric	1	1	1	1	1	1
No Response	5	4	2	8	7	5
	291	291	291	291	291	291

3. Have you an established formal laboratory quality control scheme for blood parameters?

	No. of Respondents
Yes	272
No	19
Total	291

4. In your quality control scheme, do you monitor the accuracy of your measuring process?

	No. of Respondents
Yes	274
No	14
No Response	3
	291

5. In your quality control scheme, what materials do you use for establishing the accuracy of your measuring process?
(More than one response permissible.)
(a)

	Yes	No	N/A*	N/R**	Total
Selected Patient Samples	48	241	1	0	291
Commercial preparation(s)	273	17	1	0	291
Blood from donor bank	139	151	1	0	291
Staff controls	28	262	1	0	291
Selected blood from previous day	65	225	1	0	291
DHSS trial	165	125	1	0	291

* Not Applicable ** No Response

5.(b)

Other Materials	No. of Respondents
None	259
Horse/Donkey	2
Other inter-lab. trials	25
Sequestrene blood (Source not stated)	1
'Own product'	2
Dade	1
No response	1
	291

6. In your quality control scheme, do you monitor the precision of your measuring process?

	No. of Respondents
Yes	270
No	18
No Response	3
	291

7. In your quality control scheme, what materials do you use in monitoring precision? (More than one response permissable.)

(a)

	Yes	No	N/A*	N/R**	Total
Selected patient samples	60	217	12	2	291
Commercial Preparations	164	113	12	2	291
Blood from donor bank	160	117	12	2	291
Staff controls	17	260	12	2	291
Selected blood from previous day	57	220	12	2	291
DHSS Trials	94	183	12	2	291

* Not Applicable ** No Response

(b)

Other Materials	No.Respondents
None	249
Horse/Donkey Blood	2
Other inter-lab trial	19
Sequestrene Blood (Source not stated)	1
'Own Product'	5
Dade	2
No Response	2
Not Applicable	11
	291

8. Do you participate in a commercial quality control scheme for blood counts?

	No. of Respondents
Yes	36
No	252
No Response	3
	291

9. What statistical procedures do you employ in your laboratory? (More than one response permissible.)

(a)

	No. of Respondents
Some Statistical Techniques	250
No Statistical Techniques	39
No Response	2
	291

(b)

STATISTICAL TECHNIQUES	No. of Respondents
Shewhart charts	82
Cusum charts	72
Visual inspection	163
Average of normals	73
Pragnell and Johnston (Modal control)	5
Programs on Mean and Standard Deviation	2
Average of controls	2
Patient's data	1
Modified Youden plot	1
Running means	6

10. Do you employ in your quality control scheme:

(More than one response permissible.)

	No. of Respondents
Visual method	194
Pencil and paper methods	159
Calculating machine	161
Laboratory computer	37
Computer link	8
Off-link hospital computer	1
Cumulative patient files	3
Datalab Q.C.	1
Slide rule	1

11. Do you run checks on (a) Carryover? (b) Linearity?

Percentage of Respondents.*

Laboratory Size	Twice or More/Day		Once/Day		Once/Week		Once/Month		Manufacturer's Service		Not Applicable	
	C**	L**	C	L	C	L	C	L	C	L	C	L
Fewer/Or 100	15	12	12	12	6	5	4	9	9	12	54	50
101 - 250	17	12	11	10	8	10	8	7	45	44	11	17
greater than 250	26	10	18	21	8	9	0	7	33	37	15	16
Total percentage	18	12	13	13	7	8	5	8	31	33	26	26

*Percentage of respondents minus 8 cases-carryover, 5 cases-linearity with responses other than tabulated e.g. 'when suspect', after WBC in excess of $20.00 \times 10^9/1$ etc.

** C ≡ Carryover

L ≡ Linearity

Section B: DHSS TRIALS

12. By which method(s) do you normally report results in the DHSS trials?

(More than one response permissible)

	No. Respondents
Fully Automated	203
Semi-Automated	176
Manual	149

13. How long have you been participating in the DHSS trials?

	No. Respondents
Two or More Years	257
Less than Two Years	34
	291

14. Currently, in what proportion of these trials do you participate?

	No. Respondents
Half or more	285
Less than half	3
No Response	3
	291

15. If you are unable to participate in all of the DHSS trials, is the cause most commonly due to:

	No. Respondents
Excessive workload	36
Instrument Breakdown	14
Unsuitability of Sample	13
Personnel Absence	25
Postal Delays	22
"Oversight"	1
Take part on all trials	174
No Response	6
	291

16. If you report in the fully automated method, do you use:

	No. Respondents
Coulter-S	194
Technicon	8
Other (Coulter and Technicon)	1 (=2)
Not Applicable	87
	291

17. Where possible, if there is enough blood, do you carry out duplicate determinations on the trial specimens?

	No. Respondents
Yes, Sequentially	172
Yes, Not Sequentially	61
No	57
No Response	1
	291

18. If duplicate determinations are made, do you report averages?

	No. of Respondents
Yes	145
No	85
Not Applicable	59
No Response	2
	291

19. On average, do you make the determination(s):

(More than one response permissible.)

	Yes	No	No response
Immediately after calibration and resetting	44	243	4
Immediately after full cleaning	18	269	4
Always at the same time of day	52	236	3
The day the samples arrive	272	15	4

20. Are the trial specimens analysed as:

	No. of respondents
Special samples by specified staff	29
Special samples by the day's operator	185
Unknown samples (blind)	77
	291

21. Please indicate where applicable, the most common result in the last five trials as assessed by the DHSS in the overall variance index $|R_o|$.

	Full Automated	Semi Automated	Manual
Excellent (0.0-0.5)	4	8	14
Good (0.5-1.0)	161	120	106
Acceptable (1.0-2.0)	35	34	26
Unacceptable (2.0 or more)	0	0	1
Not applicable	84	121	136
No response	7	8	18
	291	291	291

22. Do you think that these trials are being used by the DHSS to:

(More than one response permissible.)

	Yes	No	N/R*
Determine precision and accuracy of different methods	168	130	2
Determine suitability of possible reference materials	221	77	2
Determine specific laboratory capabilities	100	198	2
Collect information for use in laboratory quality control	209	89	2
Improve overall standards	283	16	1

* No Response

Questions 23, 24 and 25 refer to the results format pre-trials 77/11 and 77/12

23. For normal laboratory personnel, do you think that the presentation of trial results is:

	No. of Respondents
Easily interpreted	82
Fairly difficult to interpret	185
Very difficult to interpret	29
No Response	4
	300

24. Have you ever requested help in interpreting the trial results?

(a)

	No. of Respondents
Yes	51
No	247
No Response	2

If so, please indicate from whom:

(b)

Amateur Statistician	3
DHSS	29
Senior Staff	1
Colleague	8
Statistician	7
Coster and Lewis (Book) (see ref. Ward and Lewis (1975))	2
No response	1

24.(c)

If so, how?

Verbally	6
Telephone	20
Letter	6
Personal talk at meeting	4
No response	15

25. Would you like to see the presentation of the results changed?

(a)

	No. of respondents
Yes	132
No	155
No response	13

If so, state how (briefly)

(b)

	No. of respondents
Simplified	33
Briefer	17
Simplified with better documentation	3
Graphical (Youden Plots etc.)	15
As in revised format	10
As in Birmingham/Wellcome	10
'Range for each parameter'	10
No response	34
	132

26. Is the revised format (Trials 77/11 and 77/12) an improvement?

	No. of respondents
Yes	171
No	105
No response	24
	300

27. Do you find the blood count results from (IQC-Form 3) satisfactory?

	No. of respondents
Yes	260
No	26
No response	14
	300

28. Do you think the DHSS trials are:

	No. of respondents
Too frequent	9
Just right	224
Not frequent enough	65
No response	2
	300

29. Do you think the DHSS trials should be expanded in content?

(a)

	No. of respondents
Yes	162
No	129
No response	9
	300

29.(b)

If so, what parameters should be included?

	Frequency
Platelets	133
Serum vitamin B_{12}	32
Serum folate	32
Abnormal haemoglobins	20
Peripheral blood films	20
Reticulocytes	15
Differential counts	14
Others*	44

* This classification included some twenty five clinical parameters e.g. ESR, $R1A_3$, LAP, Bone Marrow, Plasma Viscosity.....

30. Do you use the results of the DHSS trials in any way in your own quality control?

	No. of respondents
Yes	133
No	156
	291

31. (i). Do you participate in any other kind of interlaboratory trials?

	No. of respondents
Yes	72
No	213
No response	6
	291

Type of trial	Frequency
Local	29
Regional	56

31. (ii). Do you participate in any other kind of interlaboratory trials?

(More than one response permissible).

	No. of respondents
Yes	149
No	140
No response	2
	291

Type of trial	Frequency
Local	28
Regional	69
National	79
International	30

32. Is the profession of the person(s) completing this questionnaire:

	No. of respondents
Technical	167
Medical	26
Technical and Medical	96
Scientific	1
No response	1

33. How do you think your staff react to the DHSS trials? (Indicate any differences between the reaction of medical and technical staff.)

	Medical	Technical
Guide to Lab. performance	20	20
No reaction	25	30
Interest	45	46
Favourable	51	54
Co-operative	25	34
Realisation of need	17	19
Acceptance	22	22
Usefulness invalidated by long delay in feed back	14	17
No response	24	22
Other specific reactions	48	27

34. How does your staff react to a series of 'bad' trial results?

	No. of labs
Senior staff investigate	7
No experience of this	147
Disappointment/resolve to improve	35
Concern	44
No help - feedback too slow	15
No response	26
Other specific comments	14

Other observations or comments concerning the DHSS trials.

	No. of Respondents
Difficulty in understanding Variance Index	17
Feedback too slow to be of help	132
Forms should be in same order as Coulter (i.e. Clinical parameters)	9
DHSS should send regular reference materials to cut out costly commercial preparations	4
Information concerning aims and handling procedures should be sent.	8
More time to return results (these mainly appeared to come from BFPO laboratories)	6
Reported results should be to the same number of decimal places as the reporting machines	8
Graphical Presentation/Youden Plots (These tended not to be the same as requested graphical presentation - table XXV(b)	8
No response (number of laboratories)	149
Other comments.	

7 Conclusions, Proposals and the Future.

We have discussed problems occuring currently in haematology laboratories. In order to elucidate and solve these selected, inter-related problems, we have employed a substantial element of statistical theory. In the introductory chapter, we described briefly the developments in statistics and also the developments in computing machines. A parallel description was given for some of the more important developments in the medical field. At no time did we attempt to define what we meant by statistics, rather we decided to discuss and solve the chosen problems using techniques which statisticians employ. Then it should be of interest to discuss in retrospect what constituted the statistical element. Let us attempt to define what is meant by statistics.

In the past, statistics has tended to be described as a science, concerning the experimentation, collection and presentation of data. Although these definitions do contain some truth, they tend to give the impression that statistics is somehow isolated from the real world. Thus it becomes the prerogative or pastime of academics. Kendall and Stuart (B23) have attempted to relate statistics to the real world.

'Statistics is the branch of scientific methods which deals with the data obtained by counting or measuring the properties of populations of natural phenomena. In this definition 'natural phenomen' includes all the happenings of the external world, whether human or not'.

I would disagree that statistics should be considered as a purist, academic pursuit. In fact, until quite recently many of the accepted statistical techniques were developed by experimenalists and researchers, not mathematicians. Barnett (B5) has given a less formal definition, defining statistics as 'the study of how information should be

employed to reflect on, and give guidance for action in, a practical situation involving uncertainty'.

Barnett suggests that it is perhaps better to say too little in defining statistics than too much, hoping that his writings will fill in any gaps in the definition. In the same way, we felt that it was better to show the 'what' and 'how' rather than become involved in a discussion of semantics.

Yule and Kendall (B38) have described the reaction of the layman to statistics in the following way.

'The layman's attitude towards statistics is admirably summed up in the remark that mankind is divided into two parts, those who say that figures can prove anything and those who assert that they can prove nothing. It must be admitted that this attitude is not unreasonable. From the advertising hoarding, from the electioneering platform, from the partisan press, and from a dozen other sources, the man in the street is bombarded with tendentious figures put forward to support some ex parte statement. Sometimes such figures are justifiably used to form a basis for the arguments which are built upon them; more often they give a specious picture of the truth, which may be due to ignorance or inadvertence, but has also been known to be occasioned by a deliberate wish to mislead. The layman is well aware of this fact. His attitude in distrusting all arguments based on figures is that of a reasonable man, who has not the training to distinguish for himself the true from the false, and is therefore inclined to suspect everything'.

We believe that if the statistician takes time to understand the practical problems of the 'layman' and does not blindly suggest standard statistical procedures out of hand, the 'layman' has much more belief in the results and may well see more clearly the

relevance of statistics in general. Further, it may well motivate the 'layman' to appreciate areas in which statistics can help and ultimately lead to a better understanding of the 'layman's' problem.

We have discussed problems in the 'practical situation' of the haematology laboratory, specifically with reference to peripheral blood counting. The discussion has shown that statistics is an integral and necessary part of the technician's and the clinician's knowledge. Hopefully, the staff of the haematology laboratory do not have the 'attitude of the layman' described by Yule and Kendall. We are not advocating the imposition of statistics, but rather, through education and example showing where and how statistics may help.

The problems which have been discussed have fallen into two broad classes. Firstly, the evaluation of a measuring process and secondly, the controlling of quality in the laboratory's results. Both of these classes of problems were divided into two sub-classes. Discussions relating to the evaluation of a single measuring process were treated under the heading of 'Evaluation of Machines/Methods' - Chapter 3, and 'Comparision of Different Measuring Processes' - Chapter 4. Discussions concerning the monitoring of laboratory quality were discussed under the headings of 'Intra-Laboratory Quality Control' - Chapter 5, and 'Inter-Laboratory Quality Control' - Chapter 6.

In each of the Chapters, we have shown the interrelationships of each of the topics under discussion with topics in the other chapters. It is important to bear in mind these interrealtionships, thus treating the clinical laboratory and the measuring process as dynamic systems, not some vast pre-set automata.

In Chapter 2, we presented the statistical notation and basic statistical background for the other chpaters. Further, some examples

of the effects of certain data handling procedures on clinical measurements was reported. We showed that in certain situations, when the rounding loss was similar to the precision, that effects could be large. However, when we considered the rounding within the Coulter-S, the effects were not significant. The main example concerned with treatment of replicate, specifically three replicate, measurements. It was noted that certain common laboratory procedures, e.g. the average of the two closest of three, did not necessarily give an accurate estimation of the clinical situation. Furthermore, the impression of the precision of the measuring process was usually falsely high. A number of different procedures were presented and discussed. Their behaviour was investigated by simulating a clinical measuring process and employing the procedures. As we stated, there was no single procedure which was 'best' in every possible clinical situation. For practical purposes, in situations in which infrequent but large contamination was expected, the mean of three, unless the two extreme observations differed by more than a specified amount whence the mean of the closest two was used, was found to give a reasonable procedure. This does not mean to say that this is the 'best' procedure. We note that with a back-up computer or a computer handling the clinical measurements, a more complex and perhaps better procedure could be adopted, but the benefits of such complex and arithmetically exacting schemes is lost if they have to be implemented by laboratory staff. Our approach to the problem of outliers has been one of constructing procedures which were not too severely affected by their presence. The alternative approach of identifying and subsequently discarding outliners was rejected. The cost in terms of time and manpower would probably be too great in the routine laboratory, whereas we investigated procedures for accepting all data

and handling it in such a way that the effects of outliers was minimised.

In Chapter 3, we described the various components of the haematological measuring process, with special reference to peripheral blood counting. Many sources of variability were identified, e.g. collection of specimens, administration, reagents, methodology in measuring etc. However, we confined our attention to the sources which are attributed to the measuring laboratory. This does not mean to say that the other sources are not important.

There followed an investigation of the evaluation of a single measuring machine. (Many of the ideas could be transferred to situations in which we evaluate any specific measuring methodology). This evaluation was in 'isolation'. That is to say it was an examination of the inherent characteristics of the process, using, perhaps, only standard/reference proparations. The evaluation had two parts, the physcial and the statistical evaluation. We considered that the physcial evaluation of the machine had already taken place. This evaluation concerns the costs, rate of handling of specimens, principle of measurement, supplied standards, space, servicing, requirements of staff, specimen requirements, reagents, interation of staff with device etc. Having carried out this evaluation, and having accepted the machine, we turned to the statistical evaluation of the machine. The discussion specifically concerned the Coulter-S blood cell counter. Many of the existing reports of such evaluations, describe ad hoc procedures for such evaluations. Further, many of the indices which are the basis of such evaluations, involve unjustified or inconsistent statistical assumptions. We proposed that accuracy, precision and carryover, should be assessed simultaneously and should be corrected for drift (linear). (In some situations i.e.

coded levels, it is possible to investigate linearity). In order to achieve this, we constructed two statistically equivalent designs using three different test materials, which could be either quality control blood or commercial preparations.

For practical reasons, it was felt that the Triple Design should be used in practice. However, we could not extend the construction of the Triple Design for use with more than three test materials. The other type of design, the Double Design, could easily be extended, and used in situations in which more than three test materials were employed.

The practical experiments carried out, confirmed the belief that such designs were viable in the laboratory and even for the smallest size of design gave reasonable results. (In practice with large evaluations, the basic design reported would be replicated).

The evaluation of a measuring process in 'isolation', although important cannot be considered as a complete evaluation in itself. There must be comparison with existing measuring processes for the same clinical parameter(s). This was the subject of discussion in chapter 4.

Firstly we examined some indices which may be used in the comparisons. That is, some numerical value which indicates the association between measurements or the possible diagnosis for different measuring methods. We indicated that frequently the product moment correlation coefficient could not be relied on to give an accurate indication of the true situation. From the examples, we suggest that this index is only a very rough indication of association. We recommend that in all situations a scatter diagram should be constructed. Scrutiny of such a graph often indicates non-linearity, changes in precision or some other problem.

The use of a single number to indicate, we note that it does not measure, the degree of association between two or more measuring processes is almost certainly of little use. The relationship between the measurements from the processes is too complex to be encompassed by a single index.

In turning our attention to examining the relationship between two (or more) measuring processes, we reject the use of simple linear regression. It is inappropriate as it requires one of the variables to be measured without error. Several methods of analysis which acknowledges that both sets of results have measurement error were presented. Depending on the specific situation, e.g. replication or not, more appropriate analyses could be constructed.

Finally in this chapter, we discussed the use of a hybrid experiment employing a Replicate and a Grubb's design. This hybridisation was an attempt to overcome the problem of small volume samples in haematology and the need for estimates of precision and relative bias. This scheme can only be used if we can assume that no proportional error exists.

Throughout this chapter, when discussing comparisons we meant the comparision of equivalent methods. These methods were those which could be considered to measure exactly the same constituents while purporting to measure a specific clinical parameter. We did indicate that in certain circumstances, non-equivalent methods could be investigated by transforming them into a comparison of equivalent machines. There are some situations which may not admit of this, and we remark that investigations into such situations and the appropriate analyses should be undertaken in the future.

Chapters 3 and 4 dealt with the evaluation of a measuring process, both the inherent characteristics and also the comparison

with existing measuring processes. Having investigated how we might investigate a single measuring process, we turned our attention to the laboratory as a whole and discussed methods for controlling the quality of work carried out by the laboratory. The quality of clinical work could be assessed in 'isolation' i.e. compared with past capabilities of the laboratory - Intra-laboratory quality control, and also assessed with respect to other independent laboratories - Inter-laboratory quality control. Reasons for both types of monitoring were discussed.

In chapter 5, we discussed Intra-laboratory quality control. As a preliminary the concept and construction of 'normal ranges' was discussed. The discussion of the construction and use of 'normal ranges' specifically highlighted one area of interface of the clinician with the measuring process. There are a plethora of construction methods; the interpretation of the intervals depend on the specific methods of construction used. No construction procedure was better than another. (Although we gave an example in which the use of a specific interval was inappropriate due to one of the assumptions on which the procedure was based being unjustified). The most important point to realise is that the assumptions concerning the intervals and their construction dictate the interpreation given to the 'normal range'.

A single Intra-laboratory quality control scheme cannot be given for all laboratories. The size of the laboratory, the measuring process used and the requirements of the clinical staff should all affect the quality control scheme. However, we discussed some basic principles which should apply to all schemes. Having discussed the frequency of control and graphical methods which may be employed in quality control we investigated the use of patient data in quality control. We came to the conclusion that formal methods based on

patient data would not give the control required of a routine haematology laboratory by themselves. Furthermore, we suggest that procedures using patient specimens are important in a quality control context, but the resulting data should be used to give an indication of possible problems, not as a basis for sophisticated statistical analyses. Thus the use of patient specimen results contribute to the overall 'picture' of control.

The final part of chapter 5 was given to a description of a quality control scheme which is currently in use. We suggest that the scheme presented is of a form which allows for reasonable control while allowing for the routine work of the laboratory to be carried out. The quality control described covered control of reagents and the monitoring of the laboratory from day to day and also within days.

We feel that the 'evaluation designs' of chapter 3, the Triple and Double designs may well prove of use in quality control schemes. That is they could be incorporated into existing schemes. Their frequency of use would depend on the situation under consideration, but I doubt if they could be used more than three times per day. However, if used, some computing facilities - computer link, laboratory computer or programable calculator, are required.

In chapter 6, we discussed the need and construction of inter-laboratory quality control schemes. We consider that such schemes are vital to the modern clinical laboratory, not only in helping individual laboratories, but in collecting information on methods and materials. Once again the problems concerned with materials, and the monitoring of the situation were discussed before graphical procedures were discussed. A specific inter-laboratory scheme of special interest to haematology laboratories was examined - The DHSS (LDAG) and BCSH

Quality Control Trial. Some doubt was cast on the usefulness to the participant laboratories of the variance indicies.

Finally we presented the results of a survey on the participants of the DHSS (LDAG) and BCSH haematology Quality Control Trial. The questionnaire was designed to investigate the laboratory practice the procedure employed with test materials and the understanding of the laboratory staff with respect to these trials. The majority of laboratories reported carrying out some quality control - many on accuracy with fewer reporting frequent monitoring of carryover and linearity. The most disturbing finding was that the majority of laboratories made replicate determinations with many reporting averages. This is contrary to requested procedure and will tend to give a better impression of laboratory practice than actually exists.

This point illustrates the basic problems relating to inter-laboratory trials. Basically, since the majority of measuring devices are well known and can be made to produce acceptable results, and also the majority of technical staff have a level of competence to allow for their results to be acceptable, the inter-laboratory assessments should reflect routine laboratory quality. That is the quality of the routine specimens which passes through the measuring process. However, it is reasonable for the participant laboratories to behave in a clinically acceptable way. That is if a specimen is important, extra precausions are taken e.g. cleaning of machine, preceding check of calibration, double measurements etc. However, such 'care' is counter-productive to the general philosophy of the type of trial outlined. Peculiarly, although the majority of laboratories found at least some difficulty in interpreting results, few requested help. It must be stated that although there is an onus on the administrators of any such trial to ensure a format for participants to understand,

the participants must in turn report any difficulties they experience. Many participants thought that the scheme should cover more clinical parameters although many did not want it to be more frequent. The main criticism, was the long feedback time which as far as the participants were concerned nullified any benefit of participating.

We felt that the majority of participants appreciated the need for such an inter-laboratory trial. The following proposals, we feel would improve the trial from both an overall view-point and a participant's view-point. Firstly, every participant should be made clear of the aims and objectives of the trials in general and each specific trial, if necessary, in particular. Secondly, each participant shoud be sent regular reminders of the exact trial procedures. Both suggestions could be carried out through a circulated leaflet and/or by holding regional symposia. Thirdly, the presentation of the trial results should be changed. We suggest a more graphical presentation based on a Youden plot. However, each participant should be directed to sketch the 'Ratio test' graph, i.e. the specific laboratory results against the national means. The Variance indices should be abandoned, although the raw data on which they are based should still be reported. Further, we suggest a cumulative report, say the last five trial results, shoud be presented with each participant's results.

These suggestions we see as part of a future development. However, there are other areas which should be developed and extended. We have discussed the comparison of non-equivalent measuring processes. With the increase of radio-isotopic methodologies, which are replacing some of the hitherto standard microbiological assays, a much more critical appraisal of the published comparisons should take place. In such non-equivalent comparisons the most frequent analysis is by regression.

The loss function approach in comparing measuring processes received an ultimate dismissal. This was due to the difficulty in constructing a suitable loss function for specific clinical situations. Such investigations bring us in contact with ideas frequently encountered in operations research. To construct a meaningful loss function for correct and wrong diagonosis will almost certainly lead us to consider a cost benefit analysis. The basic problem with such an analysis relate to the subjective 'weighing' of costs to sepcific patients, to patients in general, to the clinical staff, to the hospital and to the National Health Service. Thus, although a measuring methodology is 'better' than any existing methodology in terms of its statistical characteristics, the degree of reliability is not warranted by the clinician, or, unfortunately it may cost too much in relation to other parts of the necessary clinical budget.

In the area of Intra-laboratory quality control, many interesting problems exist. With the current trends in the development of micro-processors, 'a computer in each laboratory' is a reality. This may lead to greater facilities for recording easily accessed patient health records, containing cumulative haematological information. Further, such developments in the use of micro-processors allow for more and greater sophistication in quality control data handling. The new generation of automatic machines will have in-built computing facilities. Such facilities are already obtainable, see SMAC (Technicon). These automatic facilities enable drift corrections to be made. Developments may bring automatic carryover correction.

We have considered internal laboratory quality control almost totally from the point of peripheral blood counting, although the points made do carry over to other haematological investigations. However, more specific considerations should be made of the many

other haematological determinations which have become routine in the clinical laboratory. e.g. radio-isotopic measuring processes, peripheral blood films, differential counts etc.

In the area in interlaboratory quality control, we have already considered some future developments in the discussion above. The obvious areas for development relate to the extension of the clinical parameters covered by the interlaboratory trials. However, the overwhelming problem remains; and that concerns the availability of suitable stable standards.

APPENDIX 1

```
C
C     THIS PROGRAM SIMULATES AND COMPARES THE RESULTS OF
C     PROCEDURES FOR ESTIMATION OF CLINICAL PARAMETERS IN THE
C     PRESENCE OF OUTLIERS.  SIX PROCEDURES OF HANDLING TRIPLE
C     REPLICATE DETERMINATIONS ARE USED AND COMPARED WITH THE
C     MEAN OF TWO REPLICATES AND ALSO A SCHEME WHERE A
C     MAXIMUM OF THREE TRIPLE REPLICATES MAY BE REQUIRED.
C
C
C     THE MEASUREMENTS ARE SIMULATED STANDARD NORMAL VARIATES,
C     OBTAINED FROM A 'RAND' FUNCTION:  G05ADF
C
C
C     THE SIMULATION USES THE CONSTANT BIAS MODEL, SEE SECTION
C     THE BIAS RANGE IS 1,2,3,4(STANDARD DEVIATIONS, THE FILTER
C     SIZE IS 1,2,3(STANDARD DEVIATIONS) AND THE CONTAMINATION
C     RATES ARE 0,5,10,20 PERCENT.
C
C
      REAL*8 AD,BL,CB
      DIMENSION XX(8,8)
C
C     SET CONTAMINATION RATE
C
      DO 777 IC=1,21,5
      ICC=IC-1
      CONTRA=ICC/100.0
      WRITE(6,1) ICC
1     FORMAT(/'  CONTAMINATION RATE - ',I4,' PERCENT')
C
C     SET BIAS
C
      DO 666 IB=1,4
      BIAS=IB
      WRITE(6,2) IB
2     FORMAT(//10X,'BIAS -',I4,' STANDARD DEVIATIONS')
C
C      SET FILTER SIZE
C
      DO 999 IF=1,3
      DFILT=IF
      WRITE(6,3) IF
3     FORMAT(//10X' FILTER SIZE',I4,' STANDARD DEVIATIONS')
C
C     ZERO THE ACCUMULATOR MATRIX XX
C
      DO 99 I=1,8
      DO 88 J=1,8
      XX(I,J)=0.0
88    CONTINUE
99    CONTINUE
C
C     SET LOOP TO COUNT AND GENERATE 100 SETS OF TRIPLICATES
C
      DO 77 ICOUNT=1,100
      CALL XSAM(BIAS,CONTRA,A,B,C)
      X1=ONE(A,B,C)
      XX(1,1)=XX(1,1)+X1
      XX(2,1)=XX(2,1)+X1**2
      X2=TWO(A,B,C,DFILT)
      XX(1,2)=XX(1,2)+X2
      XX(2,2)=XX(2,2)+X2**2
      X3=THREE(A,B,C)
      XX(1,3)=XX(1,3)+X3
      XX(2,3)=XX(2,3)+X3**2
      X4=FOUR(A,B,C)
      XX(1,4)=XX(1,4)+X4
      XX(2,4)=XX(2,4)+X4**2
      X5=FIV(A,B,C)
      XX(1,5)=XX(1,5)+X5
      XX(2,5)=XX(2,5)+X5**2
      X6=SIX(A,B,C,DFILT)
      XX(1,6)=XX(1,6)+X6
      XX(2,6)=XX(2,6)+X6**2
      X7=SEV(A,B,C,DFILT)
      XX(1,7)=XX(1,7)+X7
      XX(2,7)=XX(2,7)+X7**2
      X8=ELGT(A,B,C,DFILT,BIAS,CONTRA)
      XX(1,8)=XX(1,8)+X8
      XX(2,8)=XX(2,8)+X8**2
77    CONTINUE
C
C     COMPUTE MEAN, STANDARD DEVIATION AND MEAN ERROR
C     FOR EACH PROCEDURE
C
      DO 66 I=1,8
      XS=XX(2,I)/99.0
      XX(3,I)=SQRT(XS)
      XMN=XX(1,I)
      XSS=XS-XMN*XMN/9900.0
      XX(2,I)=SQRT(XSS)
      XX(1,I)=XX(1,I)/100.0
66      CONTINUE
```

APPENDIX 1

```
C
C     PRINT RESULTS FOR EACH LOOP
C
      WRITE(6,4)
4     FORMAT(//10X,4X,'Y1',8X,'Y2',8X,'Y3',8X,'Y4',8X,
     *'Y5',8X,'Y6',8X,'Y7',8X,'Y8'/)
      WRITE(6,5)(XX(1,J),J=1,8)
5     FORMAT(2X,'MEAN',4X,F6.3,7(4X,F6.3))
      WRITE(6,6)(XX(2,J),J=1,8)
6     FORMAT(2X,'ST.DEV',2X,F6.3,7(4X,F6.3))
      WRITE(6,7)(XX(3,J),J=1,8)
7     FORMAT(2X,'TOTERR',2X,F6.3,7(4X,F6.3))
999   CONTINUE
888   CONTINUE
777   CONTINUE
      STOP
      END
      SUBROUTINE XSAM(BIAS,CONTRA,X1,X2,X3)
C
C     THIS SUBROUTINE FURNISHES THREE RANDOM STANDARD NORMAL
C     DEVIATES, WHICH HAVE A BIAS OF 'BIAS' AND A CONTAMINATION
C     RATE OF 100*CONTRA PERCENT
C
C     TWO 'NAG' FUNCTION ARE CALLED,
C          G05ADF - TO GENERATE THE STANDARD NORMALS
C          G05AAF - TO GENERATE A UNIFORM VARIATE
C
C
      REAL*8 AD,BD,CD,UD,VD,WD,XY
      AD=G05ADF(XY)
      BD=G05ADF(XY)
      CD=G05ADF(XY)
      UD=G05AAF(XY)
      VD=G05AAF(XY)
      WD=G05AAF(XY)
      X1=SNGL(AD)
      X2=SNGL(BD)
      X3=SNGL(CD)
      U1=SNGL(UD)
      U2=SNGL(VD)
      U3=SNGL(WD)
      IF(U1-CONTRA) 1,1,2
1     X1=X1+BIAS
2     IF(U2-CONTRA) 3,3,4
3     X2=X2+BIAS
4     IF(U3-CONTRA) 5,5,6
5     X3=X3+BIAS
6     CONTINUE
      RETURN
      END
C
C
      SUBROUTINE XORDER(A,B,C,XA,XB,XC)
C
C      THIS ROUTINE ACCEPTS REALS A,B,C AND ORDERS
C     THEM IN ASCENDING ORDER OF MAGNITUDE - XA,XB,XC.
C
      IF(A-B) 1,1,2
2     XA=B
      XB=A
      GOTO 3
1     XA=A
      XB=B
3     IF(XB-C) 4,4,5
4     XC=C
      RETURN
5     IF(XA-C) 6,6,7
6     XC=XB
      XB=C
      RETURN
7      XC=XB
      XB=XA
      XA=C
      RETURN
      END
C
C
      FUNCTION YONE(A,B,C)
C
C     THE VALUE YONE IS THE MEAN OF 'TWO OBSERVATIONS'
C     SINCE A,B,C, ARE RANDOM, TAKE A AND B
C
      YONE=(A+B)/2.0
      RETURN
      END
C
C
```

APPENDIX 1

```
C
      FUNCTION YONE(A,B,C)
C
C     THE VALUE YONE IS THE MEAN OF 'TWO OBSERVATIONS'
C     SINCE A,B,C, ARE RANDOM, TAKE A AND B
C
      YONE=(A+B)/2.0
      RETURN
      END
C
C
      FUNCTION YTWO(A,B,C,DFILT)
C
C     TWO MEASUREMENTS MEANED IF MOD(A-B).LE.DFILT
C     IF THIS IS NOT SATISFIED, A THIRD MEASUREMENT IS
C
C     USED AND THE CLOSEST TO OBSERVATIONS ARE MEANED
C
      DAB=ABS(A-B)
      IF(DAB-DFILT) 1,1,2
1     YTWO=(A+B)/2.0
      RETURN
2     DAC=ABS(A-C)
      DBC=ABS(B-C)
      IF(DAC-DBC) 3,3,4
3     YTWO=(A+B)/2.0
      RETURN
4      YTWO=(B+C)/2.0
      RETURN
      END
C
C
      FUNCTION YTHRE(A,B,C)
C
C     THIS RETURNS THE MEAN OF THREE MEASUREMENTS
C
      YTHRE=(A+B+C)/3.0
      RETURN
      END
C
C
      FUNCTION YFOUR(A,B,C)
C
C     THIS RETURNS THE MEDIAN MEASUREMENT
C
      CALL XORDER(A,B,C,Z1,Z2,Z3)
      YFOUR=Z2
      RETURN
      END
C
C
      FUNCTION YFIV(A,B,C)
C
C     THIS RETURNS THE MEAN OF THE CLOSEST TWO MEASUREMENTS
C
      CALL XORDER(A,B,C,Z1,Z2,Z3)
      IF((Z2-Z1)-(Z3-Z2)) 1,1,2
1     YFIV=(Z1+Z2)/2.0
      RETURN
2     YFIV=(Z2+Z3)/2.0
      RETURN
      END

C
      FUNCTION YSIX(A,B,C,DFILT)
C
C      IF THE RANGE OF A,B,C IS LESS THAN DFILT THE MEAN OF ALL
C      THREE IS REPORTED.  _THERWISE THE CLOSEST TWO ARE MEANED
C
      CALL XORDER(A,B,C,Z1,Z2,Z3)
      IF(Z3-Z1-DFILT) 1,1,2
1     YSIX=(A+B+C)/3.0
      RETURN
2      YSIX=YFIV(A,B,C)
      RETURN
      END
C
C
      FUNCTION YSEV(A,B,C,DFILT)
C
C      THIS RETURNS ANSCOMBE AND BARRON'S RESULT.
C     THE MEAN OF ALL IS RETURNED IF THE EXTREMES
C      DIFFER FROM THE MEAN BY LESS THAN DFILT
C     THE MEDIAN IF BOTH EXTREMES DIFFER FROM THE MEAN BY MORE
C     THAN DFILT IS RETURNED.  OTHERWISE THE ROGUE EXTREME
C     IS REPLACED BY Z2 + OR - D.
C
      CALL XORDER(A,B,C,Z1,Z2,Z3)
      XB=(A+B+C)/3.0
      IF((XB-Z1)-DFILT) 1,1,2
1     IF((Z3-XB)-DFILT) 3,3,2
3     YSEV=XB
      RETURN
2     D12=Z2-Z1
      D23=Z3-Z2
      IF(D12-DFILT) 4,4,5
4     IF(D23-DFILT) 6,6,7
7     YSEV=(Z1+Z2+Z2+DFILT)/3.0
      RETURN
5     IF(D23-DFILT) 8,8,9
9     YSEV=Z2
      RETURN
```

APPENDIX 1

```
    8   YSEV=(-DFILT+2*Z2+Z3)/3.0
        RETURN
    6    CONTINUE
        RETURN
        END
C
C
        FUNCTION YEIGT(A,B,C,DFILT,BIAS,CONTRA)
C.
C        TWO SETS OF MEASUREMENTS ARE RECORDED
C        FOR BOTH SETS THE MEAN OF THE CLOSEST TWO IS OBTAINED
C        IF THESE DIFFER BY LESS THAN DFILT THE MEAN OF THE TWO
C       IS REPORTED.    IF THIS IS NOT SO A FURTHER SET OF THREE
C       MEASUREMENTS IS TAKEN, THE MEAN OF THE CLOSEST TWO DETERMINED
C        AND THE MEAN OF THE CLOSEST TWO MEANS IS REPEORTED
C
        CALL XSAM(BIAS,CONTRA,YA,YB,YC)
        X=YFIV(A,B,C)
        Y=YFIV(YA,YB,YC)
        DXY=ABS(X-Y)
        IF(DXY-DFILT) 1,1,2
    1    YEIGT=(X+Y)/2.0
        RETURN
    2    CALL XSAM(BIAS,CONTRA,ZA,ZB,ZC)
        Z=YFIV(ZA,ZB,ZC)
        YEIGT=YMAC(X,Y,Z,DFILT)
        RETURN
        END
```

APPENDIX 2

```
C      THIS PROGRAM OBTAINS AN ESTIMATE OF ACCURACY, CARRYOVER,
C     DRIFT WITH CORRESPONDING STANDARD ERRORS.  IT ALSO GIVES AN
C     ESTIMATE OF PRECISION.  THE DATA REQUIRED IS IN VECTOR FORM.
C     THE FIRST ELEMENT IN INPUT VECTOR - 2 FOR DOUBLE DESIGN
C                                        - 3 FOR TRIPLE DESIGN.
C     THE SECOND ELEMENT IS IN THE RANGE 1-4, AND CODES THE
C     PARTICULAR CLINICAL PARAMETER INVESTIGATED.
C           1  -  'WBC'
C           2  -  'RBC'
C           3  -  'HB'
C           4  -  'MCV'
C     THERE FOLLOWS THE EIGHTEEN MEASUREMENTS CORRESPONDING TO ONE
C     OF THE FOLLOWING DESIGNS.....
C
C     DOUBLE : M H H L L H M L H M L M H L L H H M
C
C     TRIPLE : H M M M L L L H H H L H L M L M H M
C
C     FOUR 'NAG' SUBROUTINES ARE USED - DOUBLE PRECISION.
C
C     E04AAF, F01ACF, G02BUF, G02CHF.
C
C
      EXTERNAL FUNCTD,FUNCTT
      REAL*8 TH,ATOL,RTOL,STEPMX,ZERO,ONE,XX,YH,YM,YL,TT,SSMIN,XB,AH
      REAL*8 XM,XL,TV,AA,TF,TM,TE,TTF,PREC,XCOV,XCOVB,Z01,EPS,EX,ET
      REAL*8 TT1,TT2,TT3,DEG,DEGO,STH,SXH,SXM,SXL,SXB,SSS
      COMMON XA,XH,XM,XL,XB
      DIMENSION XX(18,5),XCOV(6,6),XCOVB(5,5),Z01(6)
      DATA ZERO/0.0/,ONE/1.0/,NOUT/6/,NINN/5/
      DATA ID1,ID2,ID3/'DOUB','LE  ','TRIP'/,IBLANK/'    '/
      DATA ICP1,ICP2,ICP3,ICP4/'WBC','RBC','HB','MCV'/
C
C     INITIALISE THE CARRYOVER: TH TO ZERO
C
C      SET ERROR AND TOLERANCES FOR 'NAG' ROUTINE CALL
C
      TH=ZERO
      ATOL=1.0E-5
      RTOL=1.0E-5
      STEPMX=1.0E-1
      MAXCAL=50
      IFAILA=0
C
C      READ DATA INTO FITH COLUMN OF MATRIX XX
C      (FORMAT IS F5.2 ONE MEASUREMENT ON A RECORD.)
C
C     SET FOURTH COLUMN OF XX WITH TIME CODES
C
C      READ IN THE DESIGN CODE AND GO TO APPROPRIATE 'NAG' CALL
C
      READ(NINN,1) IDESG,ICPAR
1     FORMAT(I1)
      READ(NINN,2) (XX(J,5), J=1,18)
2     FORMAT(F5.2)
      DO 3 J= 1,18
3     XX(J,4) = -1.9 E1 + DFLOAT(2*J)
      IF(IDESG - 2) 4,5,4
4     IF(IDESG - 3) 6,7,6
5     CONTINUE
C
C     CALL ROUTINES FOR DOUBLE DESIGN
C
      CALL E04AAF(TH,SSMIN,ATOL,RTOL,STEPMX,FUNCTD,MAXCAL,IFAILA)
      YH=XX(2,5)+XX(6,5)+XX(13,5)+XX(16,5)
      YH=YH-XX(4,5)-XX(7,5)-XX(14,5)-XX(18,5)
      YM=XX(7,5)+XX(9,5)+XX(12,5)+XX(18,5)
      YM=YM-XX(2,5)-XX(6,5)-XX(11,5)-XX(13,5)
      YL=XX(4,5)+XX(8,5)+XX(11,5)+XX(14,5)
      YL=YL-XX(6,5)-XX(9,5)-XX(12,5)-XX(16,5)
      GOTO 8
C
C     CALL ROUTINES FOR TRIPLE DESIGN
C
C
7     CONTINUE
      CALL E04AAF(TH,SSMIN,ATOL,RTOL,STEPMX,FUNCTT,MAXCAL,IFAILA)
      YH=XX(1,5)+XX(8,5)+XX(12,5)+XX(17,5)
      YH=YH-XX(2,5)-XX(11,5)-XX(13,5)-XX(18,5)
      YM=XX(2,5)+XX(14,5)+XX(16,5)+XX(18,5)
      YM=YM-XX(1,5)-XX(5,5)-XX(15,5)-XX(17,5)
      YL=XX(5,5)+XX(11,5)+XX(13,5)+XX(15,5)
      YL=YL-XX(3,5)-XX(12,5)-XX(14,5)-XX(16,5)
8     CONTINUE
C
C     THE MINIMUM SUM OF SQUARES IS SSMIN.
C     THE PARAMETER ESTIMATES ARE TH,XH,XM,XL,XB(FOUND IN COMMON).
C     FUNCT IS A SUBROUTINE SUPPLIED TO EVALUATE THE SUM OF SQUARES
C     FOR EACH VALUE OF TH.  IT IS BASED ON A 'NAG' MULTIPLE
C     REGRESSION SUBROUTINE.
C
C
```

APPENDIX 2

```
C
C     CONSTRUCT THE COVARIANCE MATRIX FOR THE ESTIMATES.
C     XCOV IS IN UPPER TRIANGULAR FORM.
C     THE ORDER IS DRIFT, CARRYOVER, H, M, L.
C
C
      TV=DFLOAT(12)
      AA=DFLOAT(1938)
      TF=DFLOAT(4)
      TW=DFLOAT(2)
      TE=DFLOAT(8)
      TTF=DFLOAT(24)
      PREC=SSMIN/TV
      XCOV(1,1)=AA/PREC
      XCOV(1,2)=TV*(XH-TW*XM+XL)/PREC
      XCOV(1,3)=TV*TH/PREC
      XCOV(1,4)=-TTF*TH/PREC
      XCOV(1,5)=TV*TH/PREC
      EX=XH*XH+XM*XM+XL*XL
      EY=XH*XM+XH*XL+XM*XL
      XCOV(2,2)=TE*(EX-EY)/PREC
      TT1=ONE-TW*TH
      TT2=TH*(ONE-TH)
      TT3=ONE-TW*TH+TW*TH*TH
      XCOV(2,3)=(YH+TF*TT1*(-TW*XH+XM+XL)+TV*XB)/PREC
      XCOV(2,4)=(YM+TF*TT1*(XH-TW*XM+XL)-TTF*XB)/PREC
      XCOV(2,5)=(YL+TF*TT1*(XH+XM-TW*XL)+TV*XB)/PREC
      DEG=(TF*TT3+TW)/PREC
      DEGC=(TW*TT2)/PREC
      XCOV(3,3)=DEG
      XCOV(3,4)=DEGC
      XCOV(3,5)=DEGC
      XCOV(4,4)=DEG
      XCOV(4,5)=DEGC
      XCOV(5,5)=DEG
C
C     SET ERROR AND TOLERANCES FOR 'NAG' INVERSION ROUTINE.
C
      EPS=1.0 E-6
      IA=6
      IB=5
      IFAILB=0
      CALL F01ACF(5,EPS,XCOV,IA,XCOVB,IB,Z01,L,IFAILB)
C
C     OBTAIN STANDARD ERROR OF ESTIMATES.
C
      STH=DSQRT(XCOV(3,2))
      SXH=DSQRT(XCOV(4,3))
      SXM=DSQRT(XCOV(5,4))
      SXL=DSQRT(XCOV(6,5))
      SXB=DSQRT(XCOV(2,1))
      SSS=DSQRT(PREC)
C
C     PRINT POINT ESTIMATES AND THE CORRESPONDING STANDARD ERRORS.
C
      IOUT2=ID2
      IOUT1=ID1
      IF(IDESG.EQ.3) IOUT1=ID3
      IOUT2=IBLANK
      IF(ICPAR.EQ.1) IOUT3=ICP1
      IF(ICPAR.EQ.2) IOUT3=ICP2
      IF(ICPAR.EQ.3) IOUT3=ICP3
      IF(ICPAR.EQ.4) IOUT3=ICP4
      WRITE(NOUT,601) IOUT1,IOUT2,IOUT3
      WRITE(NOUT,603) XH,SXH,XM,SXM,XL,SXL
      WRITE(NOUT,605) TH,STH,XB,SXB,SSS
601   FORMAT('   DESIGN:',2X,2A4//'   CLINICAL PARAMETER:',
     *2X,A4///)
603   FORMAT('   ESTIMATION',4X,'POINT ESTIMATE  STANDARD ERROR'/
     *'   HIGH',8X,F14.5,2X,F14.5/'   MEDIUM',6X,F14.5,2X,F14.5/
     *'   LOW',9X,F14.5,2X,F14.5/)
605   FORMAT('   CARRYOVER',3X,F14.6,2X,F14.5/'   DRIFT',7X
     *,F14.5,2X,F14.5/'   PRECISION',3X,F14.6)
C
C     COVARIANCE STRUCTURE : IN ROUTINE USE, THIS PART NEED
C
C     NOT BE USED.
C
      WRITE(NOUT,607)
      WRITE(NOUT,609) XCOV(2,1),XCOV(3,1),XCOV(3,2),XCOV(4,1),
     *XCOV(4,2),XCOV(4,3),XCOV(5,1),XCOV(5,2),XCOV(5,3),XCOV(5,4),
     *XCOV(6,1),XCOV(6,2),XCOV(6,3),XCOV(6,4),XCOV(6,5)
607   FORMAT(///'   COVARIANCE - DRIFT, CARRYOVER, H, M, L.')
609   FORMAT(//'   ',F12.8/' ',2(2X,F12.8)/' ',3(2X,F12.8)/
     *' ',4(2X,F12.8)/' ',5(2X,F12.8))
      GOTO613
6     WRITE(NOUT,611)
611   FORMAT('  INPUT ABORT ** CHECK DESIGN CODES')
613   CONTINUE
      STOP
      END
```

APPENDIX 2

```
      C
      C
            SUBROUTINE FUNCTD(TH,FT)
      C
      C
      C     THIS ROUTINE CALCULATES THE RESIDUAL SUM OF SQUARES,FT,
      C     BY MULTIPLE LINEAR REGRESSION FOR ANY GIVEN VALUE TH.
      C
      C      TWO 'NAG' ROUTINES ARE USED - G02BDF, TO OBTAIN THE NECESSARY
      C     INPUT REQUIREMENTS FOR THE REGRESSION ROUTINE G02CHF.
      C
      C
            REAL*8 XA,TH,FT,XH,XM,XL,XB,XBAR,STD,SSPZ,RZ,RESULT,COEF,RZINV
            REAL*8 CZ,WKZ
            COMMON XA,XH,XM,XL,XB
            DIMENSION XX(18,5),XBAR(5),STD(5),SSPZ(5,5),RZ(5,5)
            DIMENSION RESULT(13),COEF(4,3),RZINV(5,5),CZ(4,4),WKZ(5,5)
            DATA ZERO/0.0/,ONE/1.0/
      C
      C     SET DESIGN FOR DOUBLE
      C
            XX(1,1)=ZERO
            XX(2,1)=ONE-TH
            XX(3,1)=ONE
            XX(4,1)=TH
            XX(5,1)=ZERO
            XX(6,1)=ONE-TH
            XX(7,1)=TH
            XX(8,1)=ZERO
            XX(9,1)=ZERO
            XX(10,1)=ZERO
            XX(11,1)=ZERO
            XX(12,1)=ZERO
            XX(13,1)=ONE-TH
            XX(14,1)=TH
            XX(15,1)=ZERO
            XX(16,1)=ONE-TH
            XX(17,1)=ONE
            XX(18,1)=TH
      C
            XX(1,2)=ONE
            XX(2,2)=TH
            XX(3,2)=ZERO
            XX(4,2)=ZERO
            XX(5,2)=ZERO
            XX(6,2)=ZERO
            XX(7,2)=ONE-TH
            XX(8,2)=TH
            XX(9,2)=ONE-TH
            XX(10,2)=ONE
            XX(11,2)=TH
            XX(12,2)=ONE-TH
            XX(13,2)=TH
            XX(14,2)=ZERO
            XX(15,2)=ZERO
            XX(16,2)=ZERO
            XX(17,2)=ZERO
            XX(18,2)=ONE-TH
      C
            XX(1,3)=ZERO
            XX(2,3)=ZERO
            XX(3,3)=ZERO
            XX(4,3)=ONE-TH
            XX(5,3)=ONE
            XX(6,3)=TH
            XX(7,3)=ZERO
            XX(8,3)=ONE-TH
            XX(9,3)=TH
            XX(10,3)=ZERO
            XX(11,3)=ONE-TH
            XX(12,3)=TH
            XX(13,3)=ZERO
            XX(14,3)=ONE-TH
            XX(15,3)=ONE
            XX(16,3)=TH
            XX(17,3)=ZERO
            XX(18,3)=ZERO
      C
            IFAILC=0
            CALL G02BDF(18,5,XX,18,XBAR,STD,SSPZ,5,RZ,5,IFAILC)
            IFAILD=0
            CALL G02CHF(18,5,4,SSPZ,5,RZ,5,RESULT,COEF,4,RZINV,5
           *,CZ,4,WKZ,5,IFAILD)
            XH=COEF(1,1)
            XM=COEF(2,1)
            XL=COEF(3,1)
            XB=COEF(4,1)
            FT=RESULT(5)
            RETURN
            END
```

APPENDIX 2

```
      C
            SUBROUTINE FUNCTT(TH,FT)
      C
      C
      C     THIS ROUTINE IS SIMILAR TO FUNCTD BUT
      C     IS CALLED FOR TRIPLE DESIGNS
      C
      C
            REAL*8 XX,TH,FT,XH,XM,XL,XB,XBAR,STD,SSPZ,RZ,RESULT,COEF
            REAL*8RZINV,CZ,WKZ
            COMMON XA,XH,XM,XL,XB
            DIMENSION XX(18,5),XBAR(5),STD(5),SSPZ(5,5),RZ(5,5)
            DIMENSION RESULT(13),COEF(4,3),RZINV(5,5),CZ(4,4),WKZ(5,5)
            DATA ZERO/0.0/,ONE/1.0/
      C
      C     SET DESIGN FOR TRIPLE
      C
            XX(1,1)=ONE-TH
            XX(2,1)=TH
            XX(3,1)=ZERO
            XX(4,1)=ZERO
            XX(5,1)=ZERO
            XX(6,1)=ZERO
            XX(7,1)=ZERO
            XX(8,1)=ONE-TH
            XX(9,1)=ONE
            XX(10,1)=ONE
            XX(11,1)=TH
            XX(12,1)=ONE-TH
            XX(13,1)=TH
            XX(14,1)=ZERO
            XX(15,1)=ZERO
            XX(16,1)=ZERO
            XX(17,1)=ONE-TH
            XX(18,1)=TH
      C
            XX(1,2)=TH
            XX(2,2)=ONE-TH
            XX(3,2)=ONE
            XX(4,2)=ONE
            XX(5,2)=TH
            XX(6,2)=ZERO
            XX(7,2)=ZERO
            XX(8,2)=ZERO
            XX(9,2)=ZERO
            XX(10,2)=ZERO
            XX(11,2)=ZERO
            XX(12,2)=ZERO
            XX(13,2)=ZERO
            XX(14,2)=ONE-TH
            XX(15,2)=TH
            XX(16,2)=ONE-TH
            XX(17,2)=TH
             XX(18,2)=ONE-TH
      C
            XX(1,3)=ZERO
            XX(2,3)=ZERO
            XX(3,3)=ZERO
            XX(4,3)=ZERO
            XX(5,3)=ONE-TH
            XX(6,3)=ONE
            XX(7,3)=ONE
            XX(8,3)=TH
            XX(9,3)=ZERO
            XX(10,3)=ZERO
            XX(11,3)=ONE-TH
            XX(12,3)=TH
            XX(13,3)=ONE-TH
            XX(14,3)=TH
            XX(15,3)=ONE-TH
            XX(16,3)=TH
            XX(17,3)=ZERO
            XX(18,3)=ZERO
            IFAILC=0
            CALL G02CEF(18,5,XX,18,XBAR,STD,SSPZ,5,RZ,5,IFAILC)
            IFAILD=0
            CALL G02CHF(18,5,4,SSPZ,5,RZ,5,RESULT,COEF,4,RZINV,5
           *,CZ,4,WKZ,5,IFAILD)
            XH=COEF(1,1)
            XM=COEF(2,1)
            XL=COEF(3,1)
            XB=COEF(4,1)
            FT=RESULT(5)
            RETURN
            END
```

APPENDIX 3

DHSS Laboratory Number .....

Questionnaire on Haematology / DHSS Blood Trials.

The questionnaire is divided into two sections. Section A deals with questions relating to your normal laboratory practice. Section B deals with questions relating to the DHSS trials and their implications for your laboratory.

The responses to the questions will normally be made by ticking one of the listed boxes. In some questions, which have been clearly marked, several responses could be deemed necessary. Certain questions may not be applicable in your case. If this is so, please put a tick in the 'not applicable' box. The last two questions invite brief note-type responses. At the end of the questionnaire, an invitation is made for any pertinent comments or observations relating to the DHSS trials. If the space provided is not sufficient, please attatch any further comments on a separate sheet.

Section A: Laboratory Practice.

1. On average, how many routine blood count specimens does your laboratory handle on a normal weekday? .....

2. How are your 'Normal' blood ranges derived?

	Your own.	Manufacturer's	Dacie.	Other(specify).
RBC.				
WBC.				
Hb.				
MCV.				
MCHC.				
PCV/Hct.				

3. Have you an established formal laboratory quality control scheme for blood parameters? Yes. ☐ No. ☐

4. In your quality control scheme, do you monitor the accuracy of your measuring process? Yes. ☐ No. ☐

5. In your quality control scheme, what materials do you use for establishing the accuracy of your measuring process?
(Tick in more than one box if necessary.)

- Selected patient samples. ☐
- Commercial preparation(s). ☐
- Blood from donor bank. ☐
- Staff controls. ☐
- Selected blood from previous day. ☐
- DHSS trial. ☐
- Other(specify). .....

6. In your quality control scheme, do you monitor the precision of your measuring process?

- Yes. ☐
- No. ☐

7. In your quality control scheme, what materials do you use in monitoring precision?
(Tick in more than one box if necessary.)

- Selected patient samples. ☐
- Commercial preparation(s). ☐
- Blood from donor bank. ☐
- Staff controls. ☐
- Selected blood from previous day. ☐
- DHSS trial. ☐
- Other(specify). .....
- Not applicable. ☐

8. Do you participate in a commercial quality control scheme for blood counts?

- Yes. ☐
- No. ☐

9. What statistical procedures do you employ in your laboratory?
(Tick in more than one box if necessary.)

- None. ☐
- Shewhart charts. ☐
(Standard deviation plots)
- Cusum charts. ☐
- Visual inspection. ☐
- Average of normals. ☐
- Other(specify).....

10. Do you employ in your quality control scheme
(Tick in more than one box if necessary.)

- Visual methods? ☐
- Pencil and paper? ☐
- Calculating machine? ☐
- Laboratory computer? ☐
- Computer link? ☐
- Other(specify). .....

11. Do you run checks on

(a) Carryover

- More than twice a day? ☐
- Twice a day? ☐
- Once a day? ☐
- Once a week? ☐
- Once a month? ☐
- Only on manufacturer's service? ☐
- Not applicable. ☐

(b) Linearity

- More than twice a day? ☐
- Twice a day? ☐
- Once a day? ☐
- Once a week? ☐
- Once a month? ☐
- Only on manufacturer's service? ☐
- Not applicable. ☐

Section B: DHSS Trials

12. By which method(s) do you normally report results in the DHSS trials? (Tick in more than one box if necessary.)
- Fully automated. ☐
- Manual. ☐
- Semi-automated. ☐

13. How long have you been participating in the DHSS trials?
- Two or more years. ☐
- Less than two years. ☐

14. Currently, in what proportion of these trials do you participate?
- Half or more than half. ☐
- Less than half. ☐

15. If you are unable to participate in all of the DHSS trials, is the cause most commonly due to
- Excessive workload? ☐
- Instrument breakdown? ☐
- Unsuitability of sample? ☐
- Personnel absence? ☐
- Other(specify).....
- Take part in all trials. ☐

16. If you report in the fully automated method, do you use

- Coulter-S? ☐
- Technicon? ☐
- Other(specify).....
- Not applicable. ☐

17. Where possible, if there is enough blood, do you carry out duplicate determinations on the trial specimens?

- Yes, sequentially. ☐
- Yes, not sequentially. ☐
- No. ☐

18. If duplicate determinations are made, do you report averages?

- Yes. ☐
- No. ☐
- Not applicable. ☐

19. On average, do you make the determination(s) (Tick in more than one box if necessary.)

- Immediately after calibration and resetting? ☐
- Immediately after full cleaning? ☐
- Always at the same time of day? ☐
- The day the sample arrives? ☐

20. Are the trial specimens analysed as

- Special samples by specified staff? ☐
- Special samples by the day's operator? ☐
- Unknown samples (i.e. blind)? ☐

21. Please indicate where applicable, the most common result in the last five trials as assessed by the DHSS in the overall variance index /R_o/

	Fully automated	Semi-automated	Manual
Excellent (0.0-0.5).			
Good (0.5-1.0).			
Acceptable (1.0-2.0).			
Unacceptable (2.0 or more).			
Not applicable			

22. Do you think that these trials are being used by the DHSS to
(Tick in more than one box if necessary.)

- Determine precision and accuracy of different methods? ☐
- Determine suitability of possible reference materials? ☐
- Determine specific laboratory capabilities? ☐
- Collect information for use in laboratory quality control? ☐
- Improve overall standards? ☐

Questions 23,24 and 25 refer to the results format pre-trials 77-11 and 77-12.

23. For normal laboratory personnel, do you think that the presentation of trial results is

- Easily interpreted? ☐
- Fairly difficult to interpret? ☐
- Very difficult to interpret? ☐

24. Have you ever requested help in interpreting the trial results?

- Yes. ☐
- No. ☐

If so, please indicate from whom(DHSS, statistician etc.) .....

If so, how? .....

25. Would you like to see the presentation of the results changed?

- Yes. ☐
- No. ☐

If so, state how(briefly) .....

26. Is the revised format (Trials 77/11 and 77/12) an improvement?

- Yes. ☐
- No. ☐

27. Do you find the blood count results form (IQC-Form 3) satisfactory?

- Yes. ☐
- No. ☐

28. Do you think the DHSS trials are

- Too frequent? ☐
- Just right? ☐
- Not frequent enough? ☐

29. Do you think the DHSS trials should be expanded in content?

- Yes. ☐
- No. ☐

If so, what parameters should be included? .....

30. Do you use the results of the DHSS trials in any way in your own quality control?

- Yes. ☐
- No. ☐

31. (i). Do you participate in any other blood count trials? (Tick in more than one box if necessary,)

- No. ☐
- Local. ☐
- Regional(Area) ☐

(ii). Do you participate in any other kind of interlaboratory trial? (Tick in more than one box if necessary.)

- No. ☐
- Local. ☐
- Regional(Area) ☐
- National. ☐
- International. ☐

32. Is the profession of the person(s) completing this questionnaire

- Technical? ☐
- Medical? ☐
- Technical and medical? ☐
- Other(specify) .....

33. How do you think your staff react to the DHSS trials? (Indicate any differences between the reaction of medical and technical staff.)

34. How does yor staff react to a series of 'bad' trial results?

If you have any other observations or comments, please use the space below or append a further sheet.

Thank you very much for your co-operation.

APPENDIX 4

Director
Mr. W. Lutz, M.Sc.
Telephone: 031 - 667 1011
Ext: 2460

Medical Computing and Statistics Unit
Medical School
Teviot Place
Edinburgh EH8 9AG

September 1977.

Dear

RE: Questionnaire on Haematology / DHSS Blood Trials.

May I please seek your co-operation in completing the enclosed questionnaire. Your help is sought in obtaining information regarding the DHSS Haematology Trials, the use of trial results and the understanding of the aims and the validation of such trials. The results of this survey will be published.

The study has no commercial backing whatever. It is being organised as part of a research project being undertaken in this unit on the subject of instrument evaluation and quality control in haematology laboratories. The DHSS staff connected with the administration and running of the trials are neither responsible for this survey nor will they see individual returns. They have been consulted about the content of the questionnaire however, and have agreed to assist in its distribution.

STRICT ANONYMITY WILL BE GUARANTEED. In order to achieve this, the following procedure has been adopted. The questionnaire has been sent out with the routine DHSS trial report, thus the DHSS trial code will not be broken. Each laboratory has also received a stamped addressed envelope. When the questionnaire has been completed, it should be sealed in this envelope and posted. Each envelope carries your DHSS laboratory trial code number. This is to provide a means of determining which laboratories have not yet returned the questionnaire. This checking will be done in London by the DHSS staff. Once the replies have been so checked, the envelopes will be forwarded to Edinburgh unopened. Directions for the completion of the questionnaire will be found on it's first page.

The conclusions drawn from this study will not be entirely valid unless all questionnaires are returned with the requested information. I am conscious of the time taken in completing such a questionnaire, but I feel sure that the results of the survey will be of interest.

Yours sincerely

Ewan W. MacArthur.

APPENDIX 5

DHSS/BCSH HAEMATOLOGY QUALITY CONTROL ASSESSMENT SERVICE

To all participants: November, 1977

In one of our recent distributions, we included a questionnaire from Mr. E.W. McArthur of Edinburgh and a covering letter asking you to complete it. I would like to explain that this study was a private one, completely independent of the National Quality Control scheme and that the distribution was done by us as a favour of convenience. Neither the organisers of the scheme nor the DHSS have been responsible for the content of the questionnaire or its evaluation. The DHSS is particularly anxious to clarify the fact that the letter and questionnaire were distributed without consultation with them and that they are in no way responsible for the study.

S.M. Lewis

Book References.

B1 Aitchison J. and Dunsmore I.R.

Statistical Prediction Analysis

Cambridge University Press (1975)

B2 Armitage P.

Statistical Methods in Medical Research

Blackwell (1974) (3rd Printing)

B3 Armitage P.

Sequential Medical Trials

2nd edition, Blackwell (1975)

B4 Barlow R.E. and Proschan F.

Mathematical Theory of Reliability

SIAM (1965)

B5 Barnett V.

Comparative Statistical Inference

Wiley (1975)

B6 Calabro S.R.

Reliability: Principles and Practice

McGraw-Hill (1962)

B7 Castiglioni A.

History of Medicine

tr. Krumbhaar E.B. 2nd Edition,

Routledge and Kegan Paul (1947)

B8 Cramer H.

Mathematical Methods of Statistics

Princeton University Press (1974)

B9 Dacie J.V. and Lewis S.M.

Practical Haematology

5th edition, Churchill Livingstone (1975)

B10 David H.A.

Order Statistics

Wiley (1970)

B11 De Bruyn C.S. Van Dobben

Cumulative Sum Tests: Theory and Practice

No. 24 Statistical Monograph, Griffen (1968)

B12 Deming W.E.

Statistical Adjustment of Data

Wiley (1943)

B13 Eisenhart C., Hastay M.W., and Wallis W.A.

Techniques of Statistical Analysis

McGraw-Hill (1947)

B14 Fisher R.A. and Yates F.

Statistical Tables for Biological, Agricultural and Medical Research

6th Edition, Oliver and Boyd (1970)

B15 Foucault M.

The Birth of the Clinic

tr. Sheridan A.M., Tavistock (1973)

B16 Galen R.S. and Gambino S.R.

Beyond Normality: the Predictive Value and Efficiency of Medical Diagnoses

Wiley (1977)

B17 Documenta Geigy Scientific Tables

5th Edition, J.R. Geigy (1956)

B26 Mainland D.

Elementary Medical Statistics

2nd Edition , Saunders (1963)

B27 Mandel J.

The Statistical Analysis of Experimental Data

Wiley (1964)

B28 Martin H.F., Gudzinowicz B.J. and Fanger H.

Normal Values in Clinical Chemistry: a Guide to Statistical Analysis of Laboratory Data.

Marcel Dekker Inc. (1975)

B29 Neave H.R.

Statistical Tables

George Allen and Unwin (1978)

B30 Pearson K.

The Grammar of Science

2nd Edition, Adam and Charles Black (1900)

B31 Pearson E.S. and Hartly H.O.

Biometrica Tables for Statisticians

Cambridge University Press (1954)

B32 Shewhart W.A.

Economic Control of Quality of Manufactured Product

D. Van Nostrand (1931)

B33 Shryock R.H.

The Development of Modern Medicine

University Philadelphia Press (1936)

B34 Singer C.

Short History of Medicine: Introducing Medical Principles to Students and Non-Medical Readers

Oxford (1928)

B35 Wetherill G.B.

Sampling Inspection and Quality Control

2nd Edition , Chapman Hall (1977)

B36 Wilks S.S.

Mathematical Statistics

Wiley (1962)

B37 Woodward R.H. and Goldsmith P.L.

Cumulative Sum Techniques

ICI monograph No. 3, Oliver and Boyd (1967)

B38 Yule G.U. and Kendall M.G.

An Introduction to the Theory of Statistics

14th Edition, Griffen 1965

Amador E., Hsi B.P., Massod M.F. (1968)

An Evaluation of the 'Average of Normals' and related methods of Quality Control

Am.J.Clin.Path. 50, 1968, p.369-378.

Anscombe F.J. (1960)

Rejection of Outliers

Techno 2, 1960, p.123-147.

Anscombe F.J., Barron B.A. (1966)

Treatment of Outliers in a Sample of Size Three

J. Res Nat Bur St. 70B, 1966, p.141-147.

Bailly, M. (1975)

Discussion of the Options Followed in Processing Quality Control Data

Quality Control in Clinical Chemistry, Transactions of VI International Symposium, Geneva, April23/25 1975

Eds. Anido G., Van Kampen E.J., Rosalki S.B., Rubin M.

Barnard G.A. (1959)

Control Charts and Stochastic Processes

J.R. Stat. Soc B 21, 1959, p.239-271.

Barnett V.D. (1969)

Simultaneous Pairwise Linear Structural Relationships

Biometrics 25, 1969, p.129-142.

Berenson, G.S., Srinivasan S.R., Lopez-S A., Radhakrishnhmurthy B., Pargaonkar P.S., Deupree R.H. (1972)

Clinical Application of an Indirect Method for Quantitating Serum Lipoproteins

Clin Chim Acta 36, 1972, p.175-183.

Best W.R., Mason C.C., Barron S.S., Shepherd H.G. (1969)

Automated Twelve-channel Serum Screening

Med Clin N Amer. (Advances in Lab Diagnosis)

53, No.1, Jan. 1969, p.175.

Block F.E. (1978)

Measurement Error and Statistical Significance of an Independent Variable

The American Statistician 32, 1978, p.26.

Britt R.P., Bolton F.G., Cull A.C., Spray G.H. (1969)

Experience with a Simplified Method of Radio-Isotopic Assay of Serum Vitamin B_{12}

Brit J. Haem 16, 1969, p.457-464.

Broughton P.M.G., Buttolph M.A., Gowenlock A.H., Neill D.W., Skentelbery R.G. (1969)

Recommended Scheme for the Evaluation of Instruments for Automatic Analysis in the Clinical Biochemistry Laboratory

J. Clin Path 22, 1969, p.278-289.

Broughton P.M.G., Gowenlock A.H., McCormack J.J., Neill D.W. (1974)

A Revised Scheme for the Evaluation of Automatic Instruments for Use in Clinical Chemistry

Ann Clin Biochem 11, 1974, p.207-218.

Broughton P.M.G., Dawson J.B. (1970)

An Evaluation of Five Commercial Flame Photometers Suitable for the Simultaneous Determination of Sodium and Potassium

(Scientific Report No.4 Assoc Clin Biochemists)

Bull B.S. (1975)

A Statistical Approach to Quality Control

In Quality Control in Haematology, Symp. International Committee for Standardisation in Haematology

Ed S.M. Lewis, J.F. Coster, Pb. Academic Press.

Bull B.S., Elashoff R.M., Heilbron D.C., Couperus J. (1974)

A Study of Various Estimators for the Derivation of Quality Control Procedures from Patient Erythrocyte Indices

Am. J. Clin Path 61, 1974, p.473-481.

Carstairs K.C., Peters E., Kuzin E.J. (1977)

Development and Description of the 'Random Duplicates' Method of Quality Control for a Haematology Laboratory

Am. J. Clin Path 67, 1977, p.379-385

Cavill I. (1971)

Quality Control in Routine Haemoglobinometry

J. Clin Path 24, 1971, p.701-709.

Chanarin I. (1975)

Critical Appraisal of the PCV

In Quality Control in Haematology, Symp. International Committee for Standardisation in Haematology

Eds. S.M. Lewis, J.F. Coster. Academic Press

Chiu W.K. (1973)

Comments on the Economic Design of $\bar{X}$-Charts

J. Am. Stat Assoc. 68, 1973, p.919-921.

Cochran W.G. (1970)

Some Effects of Errors of Measurement on Linear Regression

6th Berkeley Symp. on Mathematical Statistics and Probability Vol 1. Eds. L.M. LeCam, J. Neyman, E.L. Scott. Univ. of California Press. 1972

Cooke W.T., Swan C.H.J., Asquith P., Melikian V., McFeely W.E. (1973)

Serum Alkaline Phosphatase and Rickets in Urban School Children

B.M.J. 1, 1973, p.324-327.

Cox D.R. (1951)

Some Systematic Experimental Designs

Biometrika 38, 1951, p.312-323

Cox D.R. (1952)

Some Recent Work on Systematic Experimental Design

J.R. Stat Soc 14B, 1952, p.211-219.

Crump S.L. (1951)

Present Status of Variance Component Analysis

Biometrics 7, 1951, p.1-16.

Daniel C. (1974)

Calibration Designs for Machines with Carry-over and Drift

W.J. Youden Memorial Address, Chemical Div. ASQC and SPES-ASA. Oct. 31, 1974.

Dixon W.J. (1953)

Processing Data for Outliers

Biometrics 9, 1953, p.74-88.

Dixon W.J. (1962)

Rejection of Observations

Contributions to Order Statistics

eds. A.E. Sarhan, B.G. Greenberg. Wiley 1962.

Dudding B.P., Jennett W.J. (1942)

Quality Control Charts

British Standards BS: 600R

Duncan A.J. (1956)

The Economic Design of $\bar{X}$-Charts Used to Maintain Current Control of a Process

J. Am. Stat Assoc 51, 1956, p.228-242.

Eisenhart C. (1947)

Effects of Rounding or Grouping Data

In Techniques of Statistical Analysis

Ed. C. Eisenhart, M.W. Hastay, W.A. Wallis.

McGraw-Hill 1947.

Eisenhart C. (1963)

Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems.

J.Res. Nat Bur Stands. C. Eng and Inst.

Vol.67C, No.2, 1963.

Elion-Gerritzen W.E. (1977)

Quality Control in Clinical Chemistry. The Two-Sample Plot and Improvement of Laboratory Performance

Am. J. Clin Path. 67, 1977, p.91-96.

Ewan W.D. (1963)

When and How to Use Cu-sum Charts

Techno 5, 1963, p.1-22.

Ewan W.D., Kemp K.W. (1960)

Sampling Inspection of Continuous Processes with no Auto-correlation Between Successive Results

Biometrika 47, 1960, p.363-380.

Fahraeus R. (1918)

Uber die Ursachen der verminderten Suspensionsstab der Blutkorperchen wahrend der Schwangerschaftilitat

Biochem Zeitschr 89, 1918, p.355.

Fahraeus R. (1921)

The suspension-stability of the blood

Acta Medica Scand. 55, 1921, p.1-228.

Fieller E.C. (1932)

The Distribution of the Index in a Normal Bivariate Population

Biometrika 24, 1932, p.428-440.

Fieller E.C. (1954)

Some Problems in Interval Estimation

J R Stat Soc 16B, 1954, p.175-185.

Folin O., Berglund H. (1922)

The Retention and Distribution of Amino-acids with Especial Reference to the Urea Formation

J Biol Chem 51, 1922, p.415.

Frankel S., Ahlvin R.C. (1967)

An Evaluation of the Number Plus Method of Quality Control

Am J Clin Path 48, 1967, p.248-250.

Gaylor D.W. (1956)

Equivalence of Two Estimates of Product Variance

J Am Stat Assoc 51, 1956, p.451-453.

Goodman L.A., Kruskal W.H. (1954)

Measures of Association for Cross Classifications

J Am Stat Assoc 49, 1954, p.732-764.

Grubbs, F.E. (1948)

On Estimating Precision of Measuring Instruments and Product Variability

J Am Stat Assoc 43, 1948, p.243-264.

Grubbs F.E. (1950)

Sample Criteria for Testing Outlying Observations

Ann Math Stat 21, 1950, p.27-58.

Grubbs F.E. (1969)

Procedures for Detecting Outlying Observations on Samples

Techno 11, 1969, p.1-20.

Grubbs F.E. (1973)

Errors of Measurement, Precision, Accuracy and the Statistical Comparison of Measuring Instruments

Techno 15, 1973, p.53-66.

Guttman I, Smith D.E. (1969)

Investigation of Rules for Dealing with Outliers in Small Samples from the Normal Distribution: I Estimation of the Mean

Techno 11, 1969, p.527-550.

Haden R.L. (1923)

Accurate Criteria for Differentiating Anaemias

Arch Int Med 31, 1923, p.766-780.

Hahn G.J., Nelson W. (1970)

A Problem in the Statistical Comparison of Measuring Devices

Techno 12, 1970, p.95-102.

Hamilton P.J. Davidson R.L. (1973)

The Interrelationships and Stability of Coulter-S Determined Blood Indices

J Clin Path 26, 1973, p.700-705.

Hanumara R.C. (1975)

Estimating Imprecisions of Measuring Instruments

Techno 17, 1975, p.299-302.

Hart B.I. (1942)

Tabulation of the Probabilities for the Ratio of the Mean Square Successive Difference to the Variance

Ann Math Stat 13, 1942, p.207-214.

Healy M.J.R. (1969)

Normal Values from a Statistical Viewpoint

Ann Clin Biochem 6, 1969, p.12

Henry R.J. (1960)

Improper Statistics Characterizing the Normal Range

Am J Clin Path 34, 1960, p.326-327.

Herrera L. (1958)

The Precision of Percentiles in Establishing Normal Limits in Medicine

J Lab Clin Med 52, 1958, p.34-42.

Hill H.M. (1960)

Experimental Designs to Adjust for Time Trends

Techno 2, 1960, p.67-82.

Hirst, A.D., Howorth P.J.N. (1975)

Estimation of Carryover in Continuous Flow Systems

Ann Clin Biochem 12, 1975, p.32.

Hoffmann R.G., Waid M.E. (1963)

The Number Plus Method of Quality Control of Laboratory Accuracy

Am J Clin Path 40, 1963, p.263-269.

Hoffmann R.G., Waid M.E. (1965)

The 'Average of Normals' method of Quality Control

Am J Clin Path 43, 1965, p.134-141.

Howe W.G. (1969)

Two-Sided Tolerance Limits for Normal Populations - some improvements

J Am Stat Assoc 64, 1969, p.610-620.

Jackson, J.E. (1959)

Quality Control Methods for Several Related Variables

Techno 1, 1959, p.359-377.

Jaech J.L. (1971)

Further Tests of Significance for Grubbs's Estimators

Biometrics 27, 1971, p.1097-1101.

Jaech J.L. (1976)

Large Sample Tests for Grubbs's Estimators of Instrument Precision with More than Two Instruments

Techno 18, 1976, p.127-133.

Johnston N.L. (1961)

A Simple Theoretical Approach to Cumulative Sum Control Charts

J Am Stat Assoc 57, 1961, p.835-840.

Jorgensen S., Warburg E.J., (1927)

The Indices and Diameters of the Erythrocytes and the Best Haematological Criterion of Pernicious Anaemia: I and II

Acta Medica Scand 66, 1927, p.109-186.

Kruskal W.H. (1960)

Some Remarks on Wild Observations

Techno 2, 1960, p.1-3.

Laessig R.H., Schwartz T.H., Paskey T.A., Indriksons A.I. (1975)

A Proposed Blind Routine Quality Control System for Multichannel Analyzers

Am J Clin Path 64, 1975, p.225-235.

Leclercq R. (1975)

Interlaboratory Quality Control using the Daily Means of Patient Results

in Quality Control in Clinical Chemistry, Transactions of the VI International Symposium, Geneva April 23/25 1975 Eds G Anido, E.J. Van Kampen, S.B. Rosalki, M. Rubin.

Lewis S.M. (1975)

Standards and Reference Preparations

in Quality Control in Haematology, Symposium of the International Committee for Standardisation in Haematology Eds. S.M. Lewis and J.F. Coster.

Lewis S.M. (1976)

Comments on Letter by Pragnell and Johnston

J Clin Path 29, 1976, p.955.

Lewis S.M., Burgess B.J. (1969)

Quality Control in Haematology: Report of Interlaboratory Trials in Britain

B.M.J. 4, 1969, p.253-256.

Lieblein J. (1952)

Properties of Certain Statistics Involving the Closest Pair in a Sample of Three Observations

J Res Nat Bur St 48, 1952, p.255-268.

Lindley D.V. (1947)

Regression Lines and Their Functional Relationship

J R Stat Soc 9, 1947, p.218.

McGilchrist C.A., Woodyer K.D. (1975)

Note on a Distribution-Free Cusum Technique

Techno 17, 1975, p.321-325.

Mainland D. (1971)

Remarks on Clinical 'Norms'

Clin Chem 17, 1971, p.267-274.

Maloney C.L., Rastogi S.C. (1970)

Significance Test for Grubbs's Estimators

Biometrics 26, 1970, p.671-676.

Mandansky A. (1959)

The Fitting of Straight Lines when Both Variables are Subject to Error

J Am Stat Assoc 54, 1959, p.173-205.

Mandel, J., Lashof T.W. (1959)

The Interlaboratory Evaluation of Testing Methods

ASTM Bulletin No 239 p.53 July

Marsaglia G. (1956)

Ratios of Normal Variables and Ratios of Sums of Uniform Variables

J Am Stat Soc 60, 1965, p.193-204.

Matthews D.M. (1962)

Observations on the Estimation of Serum Vitamin B_{12} Using Lactobacillus leichmannii

Clinical Science 22, 1962, p.101-111

Maxwell, E.A. (1974)

Estimating Variances from One or Two Measurements on Each Sample

The American Statistician 28, 1974, No.3, p.96-97.

Mitchell F.L. (1969)

The Implementation of Quality Control and Factors Affecting its Success

Ann Clin Biochem 6, 1969, p.119-122.

Morgan W.A. (1939)

A Test for the Significance of the Difference between the two variances on a sample from a Normal Bivariate Population

Biometrika 31, 1939, p.13-19.

Murphy E.A. (1972)

The Normal, and the Perils of the Sylleptic Argument

Perspect Biol Med 15, 1972, p.566-582.

Nair K.R., Shrivastava M.P. (1942)

On a Simple Method of Curve Fitting

Sankhya 6, 1942, p.121-132.

Padmore G.R.A. (1969)

Quality Control in the Small Laboratory

Ann Clin Biochem 6, 1969, p.114-118.

Page E.S. (1954)

Continuous Inspection Schemes

Biometrika 41, 1959, p.100-115.

Pinkerton P.H., Spence I., Ogilvie J.C., Ronald W.A., Marchant P., Ray P.K. (1970)

An Assessment of the Coulter Counter Model S

J Clin Path 23, 1970, p.68-76.

Pitman E.J.G. (1939)

A note on Normal Correlation

Biometrika 31, 1939, p.9-12.

Pragnell D.R., Johnston P.H. (1976)

Quality Control for Coulter Counter Model-S

J Clin Path 29, 1976, p.955.

Prangell, D.R., Johnston P.H. (1977)

A New Method of Quality Control for the Coulter Model S Counter

J Clin Path 30, 1977, p.487-491.

Prescott R.J., Hedley A.J. Paterson C. (1973)

Serum Alkaline Phosphatase and Rickets

B.M.J. 2, 1973, p.47.

Radin N. (1967)

What is Standard?

Clin Chem 13, 1967, p.55-76.

Ramsay D.M., Robertson E.P., MacArthur E.W. (1977)

Comparison of Anticoagulants for the Preservation of Prothrombin Time Specimens

J Clin Path 30, 1977, p.766-769.

Raven J.L., Robson M.B., Morgan J.O., Hoffbrand A.V. (1972)

Comparison of Three Methods for Measuring Vitamin B_{12} in Serum, Radioisotopic, Euglena gracilis and Lactobacillus leichmannii

Brit J Haem 22, 1972, p.21-31.

Reed A.H. (1972)

Misleading Correlations in Clinical Applications

Clin Chim Acta 40, 1972, p.266-268.

Rietz H.L., Mitchell H.H. (1910)

On the Metabolism Experiment as a Statistical Problem

J Biol Chem 8, 1910, p.297.

Robbins H. (1944)

On Distribution-Free Tolerance Limits in Random Sampling

Ann Math Stat 15, 1944, p.214-216.

Roberts S.W. (1959)

Control Chart Tests Based on Geometric Moving Averages

Techno 1, 1959, p.239-250.

Russell T.S., Bradley R.A. (1958)

One Way Variances in a Two Way Classification

Biometrika 45, 1958, p.111-129.

Sacker L.S. (1975)

Specimen Collection

in Quality Control in Haematology, Symposium of the International Committee for Standardisation in Haematology.

Ed. S.M. Lewis and J.F. Coster. Academic Press.

Seracci R. (1974)

The Power (Sensitivity) of Quality Control Plans in Clinical Chemistry

Am J Clin Path 62, 1974, p.399-406.

Schoen J., Brooks S.H. (1970)

Judgement Based on 95% Confidence Limits: A Statistical Dilemma Involving Multitest Screening and Proficiency Testing of Multiple Specimens

Am J Clin Path 53, 1970, p.190-193.

Searle S.R. (1971)

Topics in Variance Component Estimation

Biometrics 27, 1971, p.1-76.

Seth G.R. (1950)

On the Distribution of the Two Closest Among a Set of Three Observations

Ann Math Stat 21, 1950, p.298-301.

Sharpe A.A. (1970)

Proposed Testing Schedule for Evaluation of Equipment in Haematology Laboratories

J Clin Path 23, 1970, p.333-335.

Shukla G.K. (1973)

Some Exact Tests of Hypotheses about Grubbs' Estimators

Biometrics 29, 1973, p.373-377.

Smith H.F. (1950)

Estimating Precision of Measuring Instruments

J Am Stat Assoc 45, 1950, p.447-451.

Speckman J.A. (1964)

Determinations Based on Duplication of Readings

J Res Nat Bur St 68B, 1964, p.49-53.

Stuart A. (1953)

The Estimation and Comparison of Strengths of Association in Contingency Tables

Biometrika 40, 1953, p.105-110.

Taylor H.M. (1968)

The Economic Design of Cumulative Sum Control Charts

Techno 10, 1968, p.479-488.

Thompson W.A. (1963)

Precision of Simultaneous Measurement Procedures

J Am Stat Assoc 58, 1963, p.474-479.

Tonks D.B.A. (1963)

A Study of the Accuracy and Precision of Clinical Chemistry Determinations in 170 Canadian Laboratories

Clin Chem 9, 1963, p.217-233.

Van Assendelft O.W., Holtz A.H. (1975)

Concepts of Interlaboratory Trials: An International Haematological Survey

in Quality Control in Haematology, Symposium of the International Committee for Standardisation in Haematology

Eds S.M. Lewis and J.F. Coster. Academic Press.

Van Peenen H.J., Lindberg D.A.B. (1965)

The Limitations of Laboratory Quality Control with Reference to the 'Number Plus' Method

Am J Clin Path 44, 1965, p.322-330.

Veale J.R., Huntsberger D.V. (1969)

Estimation of a Mean when One Observation Maybe Spurious

Techno 11, 1969, p.331-339.

Ward P.G., Lewis S.M. (1975)

Inter-laboratory Trials: A National Proficiency Assessment Scheme in Britain

in Quality Control in Haematology, Symposium of the International Committee for Standardisation in Haematology

Eds S.M. Lewis, J.F. Coster. Academic Press.

Whitby L.G. (1969)

Medical Implications of Quality Control in Clinical Biochemistry

Ann Clin Biochem 6, 1969, p.104-107.

Widmark R.M., Weschler W.A. (1970)

Evaluation of a Fully Automatic Platelet Counting Technique

Technicon Quarterly 2, 1970, p.23-28.

Wilke T.A. (1966)

A Note on Contaminated Samples of Size Three

J Res Nat Bur St 70B, 1966, p.149-151.

Wilks, S.S. (1941)

Determination of Sample sizes for setting Tolerance Limits

Ann Math Stat 12, 1941, p.91-96.

Wilks S.S. (1942)

Statistical Prediction with Special Reference to the Problem of Tolerance Limits

Ann Math Stat 13, 1942, p.400-409.

Williams R.M. (1952)

Experimental Designs for Serially Correlated Observations

Biometrika 39, 1952, p.151-167.

Wintrobe M.M. (1929)

The Volume and Haemoglobin Content of the Red Blood Corpuscle

Am J Med Sci 177, 1929, p.513-523.

Wintrobe M.M. (1930)

The Erythrocyte in Man

Medicine 9, 1930, p.195-255.

Wintrobe M.M. (1931-32)

The Size and Haemoglobin Content of the Erythrocyte

J Lab Clin Med 17, 1931-32, p.899-912

Wintrobe M.M. (1933[a])

Macroscopic Examination of the Blood

Am J Med Sci 185, 1933, p.58-71.

Wintrobe M.M. (1933[b])

Blood of Normal Men and Women

Johns Hopkins Hosp Bull 53, 1933, p.118-130.

Wootten I.D.P., King E.J. (1953)

Normal Values for Blood Constituents: Inter-hospital Differences

Lancet 1, 1953, p.470-471.

Youden W.J. (1950)

Index for Rating Diagnostic Tests

Cancer 3, 1950, p.32-35.

Youden W.J. (1954)

Instrumental Drift

Science 120 1954 no. 3121, p.627

Youden W.J. (1959)

Graphic Diagnosis of Interlaboratory Test Results

Industrial Quality Control 15, 1959, No.11, p.24-28.

Youden W.J., Connor W.S., Severo N.C. (1959)

Measurements Made by Matching with Known Standards

Techno 1, 1959, p.101-109.

Zender R., Linder A. (1975)

Interlaboratory Comparison of Both Accuracy and Precision by a Two Sample Method

in Quality Control in Clinical Chemistry, Transactions of the VI International Symposium, Geneva, April 23/25 1975 Eds G. Anido, E.J. Van Kampen, S.B. Rosalki, M. Rubin.