SOME STATISTICAL INVESTIGATIONS IN HUMAN GENETICS LINKAGE.

by

MONA ABDALLA

A dissertation submitted to the

University of Glasgow

for the degree of

Doctor of Philosophy

1989

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

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor , Mr. A. D. McLaren for his advice, patience and encouragement throughout the completion of this thesis. I would also like to thank Dr. J. Yates for introducing me to the problem and for his generous assistance.

Last, but not least, I would like to thank my family, my parents and my husband for their continuous encouragement and moral support, to whom, I would like to dedicate this thesis.

Summary

The aim of this thesis is to investigate some statistical aspects of techniques developed in genetic linkage analysis.

In chapter one we provide the reader with a simplified introduction to some of the basic concepts of genetic linkage which are essential for the understanding of the work developed later, as well as giving a short summary of the relevant work published in the genetical literature.

In chapter two and three, the problem of unknown orders of three loci known to be on the same chromosome is studied. Criteria used by the geneticists to test the different orders are mentioned and then studied in various details.

In chapter four and five a comparative study between different ways of constructing interval estimates in linkage study are investigated. In chapter four certain methods of approximation for the likelihood function approach are compared under a three loci set up; on the other hand chapter five compares the likelihood and the Bayesian approach under a two loci set up.

In chapter six we study the Bayesian approach in providing a point estimate for the probability of an unborn child being at risk of carrying a genetical disease given his family pedigree. Application of this method as opposed to the likelihood approach is presented using an example from the genetical literature.

In chapter seven a discussion of the work is presented and possible extensions are suggested.

TABLE OF CONTENTS

CHAPTER ONE: Genetical background

1.1 Introduction

Although cells within a single plant or animal can vary widely in structure, shape and function, they all represent units of living material and have some important properties in common. The nuclei of any of these cells are essentially alike in term of having genes, chromosomes and other factors related to inheritance.

The gene is the unit of inheritance as it carries from generation to generation the information that specifies the characteristics of the plant or animal. Experiments had demonstrated that the nucleic acid, DNA, is the chemical of which genes are composed(Gardener 1975). The genes which are numerous, could be seen as extremely small material particles lying in certain linear order along microscopic bodies called chromosomes situated within the cell nucleous. More precisely, each gene has a certain place called *locus* on a particular chromosome.

The chromosomes occur in similar, or in *homologous*,pairs in all body cells except in the reproductive cells where they are generally single units. The number of pairs of chromosomes are usually constant for each species. In a human non reproductive cell nucleous, 23 pairs of chromosomes are present, a special pair of them controls the inheritance of sex, as well as other genetic traits, and are called sex chromosomes; the other 22 pairs are known as *autosomal* chromosomes.

Since the chromosomes occur in pairs, the loci and genes occupying them do the same. On the mean time many of these genes are *polymorphic*, (i.e) they occur in different forms or *alleles*. For example, in human genetics, the ABO blood group locus is

under the control of three alleles A^*, B^* and O^*; therefore six different genotypes, shown in column one of table(1.1), are possible. If a certain individual carries the same allele at a given gene pair he is said to be *homozygous* ((e.g) A^*A^*), and is called *heterozygous* if he carries two different alleles ((e.g)A^*O^*).

The actual appearance or expression of a particular genotype, as determined by some appropriate measurement, is called the *phenotype*, it is related to the genotype in a way that depends on the particular behaviour of the genes concerned. The different alleles of these genes could be either dominant, recessive or codominant. If for a certain diallelic gene locus, (i.e) with two different alleles H and h for example, H is completely *dominant* then individuals with genotype HH and Hh are alike phenotipically. In the heterozygous genotype Hh, h is completely masked and is called a *recessive* allele. A phenotype h would corresponds, therefore, to only one possible genotype hh. On the other hand if in a heterozygous individual Hh, H and h are fully expressed phenotipically then both alleles are *codominant*. For the ABO blood group locus, alleles A^* and B^* are codominant whereas allele O^* is recessive. Thus only four different phenotypes can be achieved for that locus as seen in table(1.1).

Virtually all normal cells can reproduce themselves. However sex or germ cells, called *gametes*, can initiate reproduction of an entire organism. When ordinary body cells divide and multiply, the cell nucleous undergoes a process of division called *mitosis*, which results in two daughter cells each having a full set of paired chromosomes exactly like the parent cell.

In the production of gametes a different mechanism called, *meiosis*, is processed, during which the chromosome number is

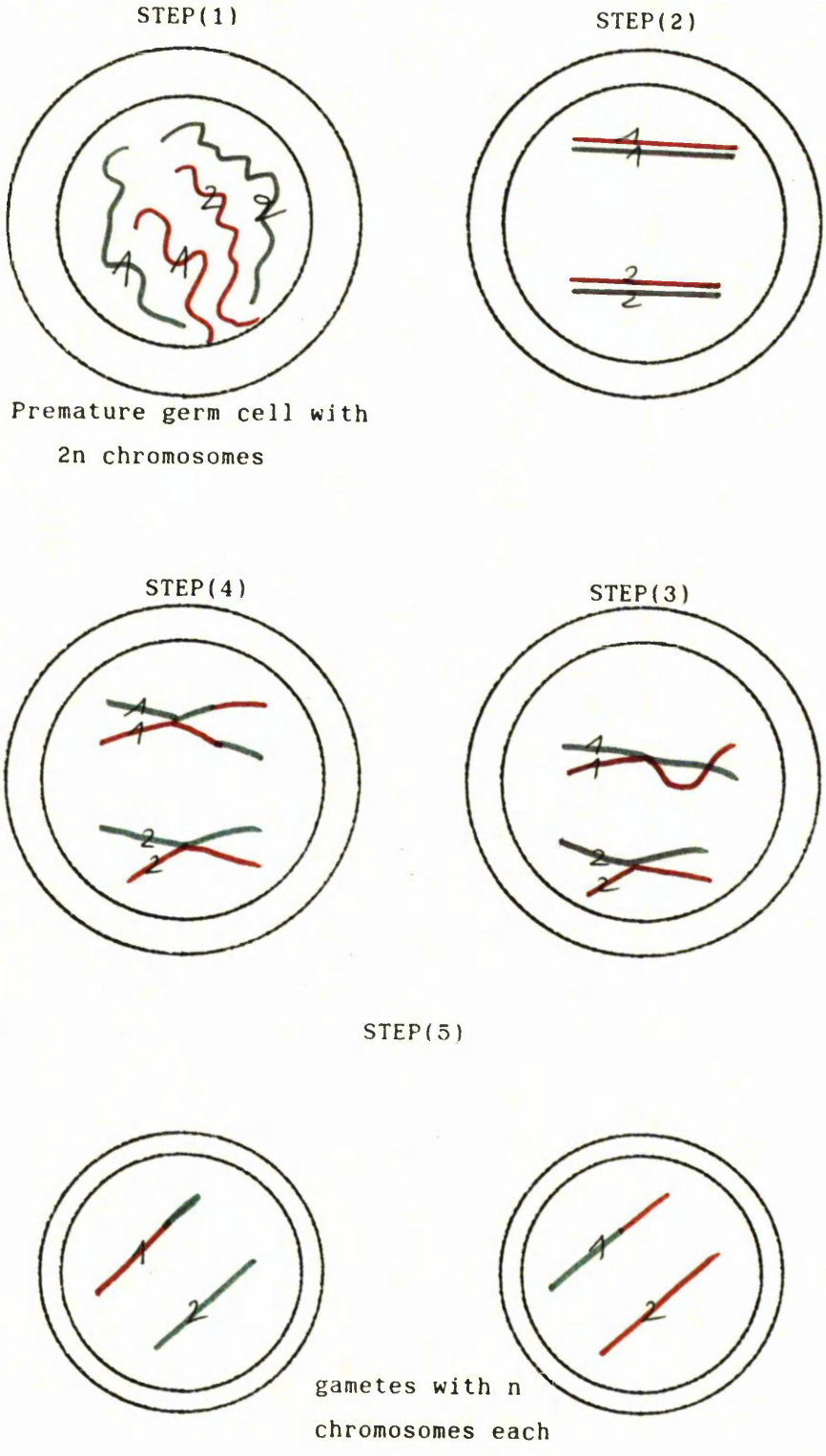
changed from the diploid number or 2n number, characteristic of

Table (1.1) Genotypes for the ABO blood group locus.

Genotype	Type of genotype	Phenotype
A^*A^*	Homozygote	A
B^*B^*	Homozygote	B
O^*O^*	Homozygote	O
A^*O^*	Heterozygote	A
B^*O^*	Heterozygote	B
A^*B^*	Heterozygote	AB

body cells and premature germ cells, to the haploid or n number which is characteristic of the gametes. Figure(1.1) shows the main simplified steps which occur during meiosis for an imaginary premature germ cell which includes only two pairs of chromosomes. In the first step we can see all chromosomes appearing singly in the nucleous of the cell, where the green chromosomes come from one parent whereas the red ones come from the other parent. During meiosis homologous chromosomes are brought together and lie side by side with corresponding loci aligned. At this stage both chromosomes will be held together at a place called the *centromere* and then they will start interchanging genetic material. Breaks may then occur at corresponding points on each chromosome, after which the chromosomes rejoin with interchange of partners, this phenomenon is called the phenomenon of *crossing-over*, which could be seen in step 3 and 4 of figure (1.1). The final step involves the division of the cell into two resulting gametes each with a single set of disimilar chromosomes. In this final step when the two different pairs of chromosomes segregate simultanously, they do so independently from each other.

Figure(1.1) A simplified plot of the division of an imaginary cell during meiosis



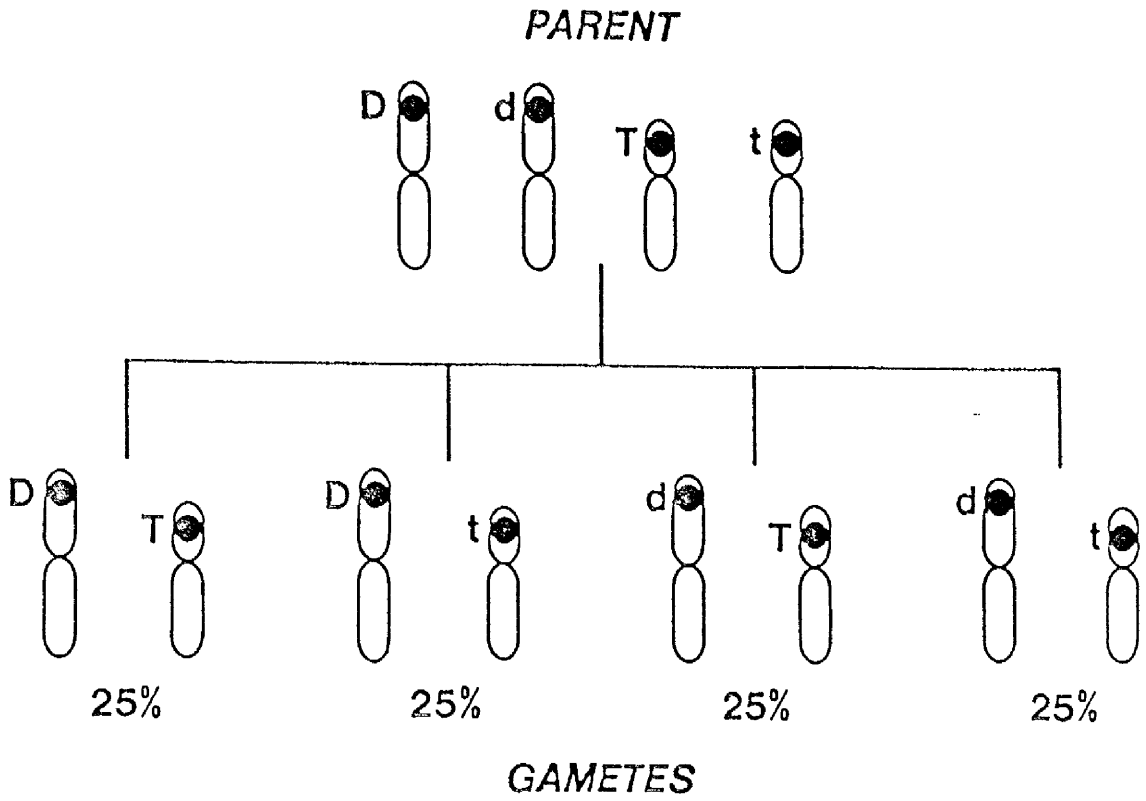
When fertilization occurs a sperm carrying a haploid number of chromosomes from the male parent is united with an ovum carrying a haploid number of chromosomes from the female parent. The fertilized egg, *zygote*, will then develops to produce an organism in each body cell of which one gene is derived from one parent and one from the other.

1.2 Linkage

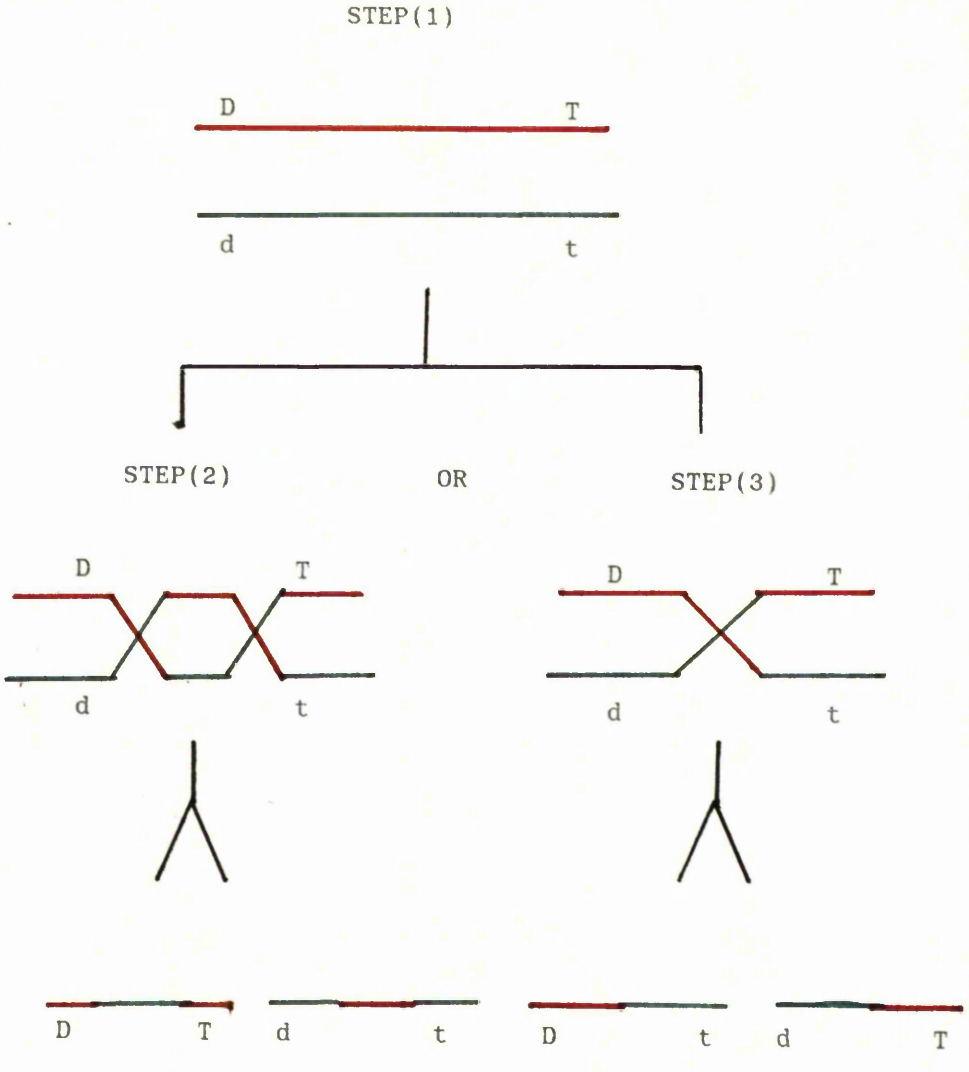
Figure(1.2) (Yates, 1986) shows two chromosome pairs, one bearing the locus for a disease caused by an abnormal gene D with the corresponding normal allele d, and the other bearing a locus, called the *marker* locus, with alleles T and t. Transmission of the disease or normal alleles into the gametes at meiosis will be independent of the transmission of the marker alleles. All four possible types of gametes are therefore equally likely. In general we can say that genes whose loci lie on different chromosomes will segregate independently. On the other hand, genes whose loci lie on the same chromosome will tend to be handed on together. This resulting disturbance of independent assortment is called the phenomenon of *linkage* and valuable information about the segregation of diseases in some families could be provided by the linked markers.

But due to the phenomenon of crossing-over, alleles at neighbouring loci will not invariably segregate together. So, if both disease and marker loci were on the same chromosome and arranged as in figure(1.3), then if there is an even number of crossing-over between the two loci, the two resulting gametes will be called *non recombinant* gametes as they are indistinguishable from the parental chromosome _step(2). On the other hand if the number of crossing-over is odd, then the two resulting gametes will be *recombinant*, (i.e) generating a new

Figure(1.2) The segregation of two loci on different chromosomes



Figure(1.3) The segregation of two loci on the same chromosome



Non-recombinant gametes

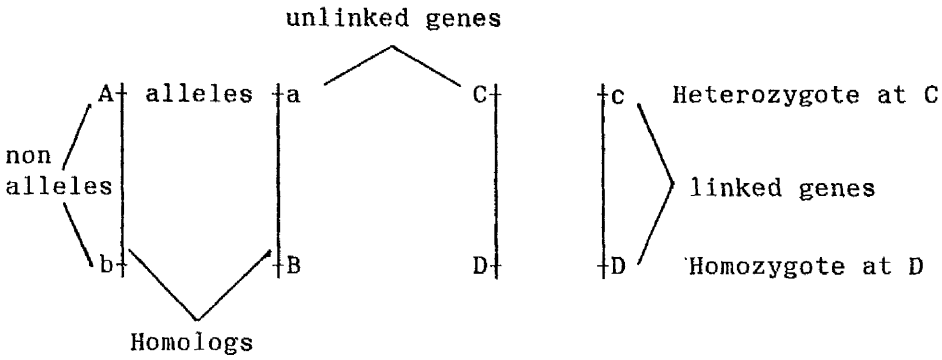
Recombinant gametes

combination Dt and dT _step(3). The number of recombinant gametes expressed as a fraction of the total number of gametes is the *recombination fraction*, denoted usually by θ. Note that, no information could have been deduced about recombination from figure(1.3) if the parental chromosomes were not doubly heterozygote at both disease and marker loci. On the mean time the *phase* of doubly heterozygote loci is essential in distinguishing a recombinant from a non recombinant gamete. In the above example the parental genotype phase was DT/dt which means that the arrangement of the alleles on the two chromosomes was as follows, D and T on one chromosome and d and t on the other. The other possible phase would have been Dt/dT, under which a Dt gamete would be a non recombinant gamete.

The extent of linkage depends on the closeness of the two loci. If they are very close, crossing-over will be rare and the number of recombinant gametes very small, hence θ near zero. The further apart the loci are the greater the recombination fraction. When the two loci are a long way apart, odd and even number of crossing_over will be equally frequent making the four possible types of gametes DT, Dt, dT and dt equally likely (i.e) θ=0.5. This case is indistinguishable from the case where the loci are on different chromosomes. Actually there is independent assortment and linkage can no longer be detected. To some extent, therefore, the recombination fraction could be used as a measure of distance between any two loci.

Some of the genetical definitions that have been used so far are summarised in figure(1.4). Table(1.2) (Yates, 1986) on the other hand summarises the relation between the recombination fraction and the position of the two loci.

Figure(1.4) A summary of some of the genetical definitions



Table(1.2) Dependence of the recombination fraction on the relative positions of two loci

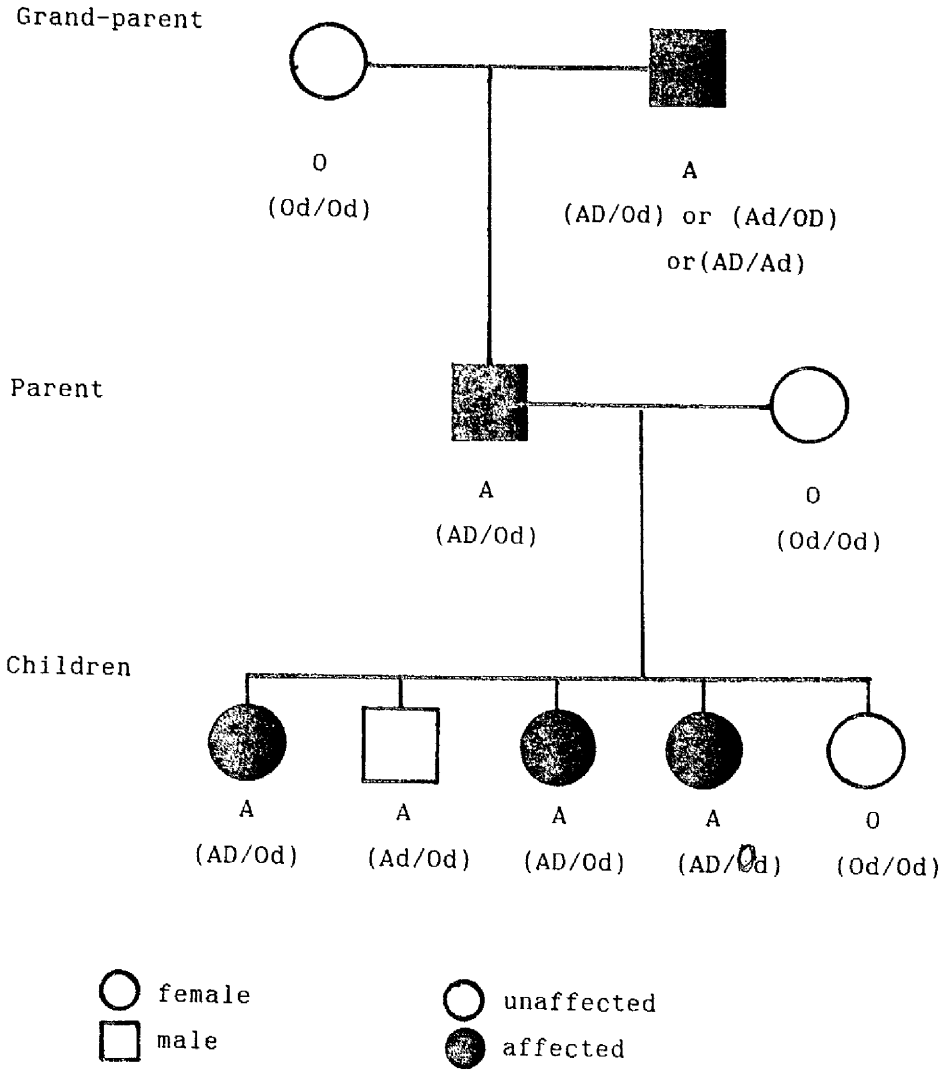
	loci on the same chromosome			loci on different chromosomes
	very close	nearby	far apart	
frequency of crossover bet. the 2 loci	rare	some	frequent	——
linkage	present	present	absent	absent
θ	0%	1—49%	50%	50%

(1.3) Marker loci

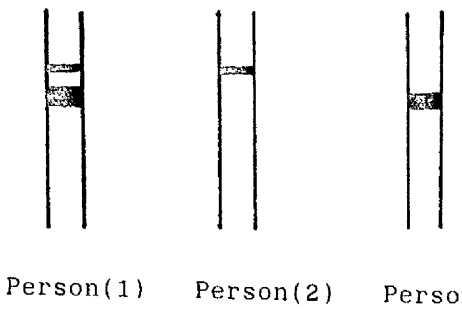
As seen above, linkage study for a disease locus has to be related to a locus of a polymorphic gene called the *marker locus*. Traditionally, examination of the DNA was not available, therefore the genotype of a person could only be inspected through his phenotype. Thus a traditional genetic marker had to be some polymorphic gene which could be observed phenotipically. The four types, or phenotypes, of blood group which are determined by a special test are a good example of such a traditional marker. Figure(1.5) shows the family tree of a certain family affected by an autosomal dominant disorder. The disease locus is determined by two alleles D, which denotes the disease allele, and d which denotes the normal one. The blood group of each individual of the family is as indicated in the figure, dark symbol on the other hand shows the affected individuals. The affected man in the second generation has received the disease allele together with the blood group A from his father and the normal allele and blood group O from his mother. If these two loci are on the same chromosome, then this individual's genotype including phase will be known for certainty as AD/Od. By analysing the offspring's genotype of this family we can see that the father had produced four non recombinant gametes, three with haplotype AD and one with haplotype Od, and one recombinant gamete with haplotype Ad. (To help the reader understanding this example the genotype of each individual which could be easily inferred from the phenotypes are written in parentheses under his or her symbol in the family tree).

These traditional markers are relatively few in number. In the past this limited the clinical usefulness of linkage, since markers close to diseases of interest could seldom be found. Now

Figure(1.5) A family tree of a family affected with a certain autosomal dominant disorder.



Figure(1.6) A schematic plot of the RFLP's on human chromosomes



the situation has changed dramatically with the development of molecular genetic techniques. At molecular level, different alleles of a gene correspond to differences or variations in the DNA sequences (Ott 1985 page19). Some of these variations are phenotipically expressed as clinical syndromes (diseases), but many others are without any classical phenotipic manifestation. A recent technique called "recombinant DNA methods" allows specialists to detect and reveal such variation in the DNA sequence between the two homologous chromosomes at a certain site, where a site is just a sequence of neighbouring loci. This method relies, firstly on some enzymes which can cut the human chromosome into small fragments according to the recognition of a specific sequence in the double stranded DNA. So that if this DNA sequence on the homologous chromosome is slightly different from the previous one, then the resulting fragments will be of different lengths. The second step of this method is how to make the difference in these fragments' lengths, known as restriction fragment length polymorphism or in short RFLP, phenotipically visible; which is a technical procedure of no interest to the present study (an exact quotation, from Ott 1985, explaining this method could be seen in appendix(1)). Nevertheless, a schematic representation of the final step after molecular analysis of the two homologous chromosomes at a certain site would be of interest to us, this is actually shown in figure(1.6). Each lane in the figure represents the genotype of a person at this site, (i.e) both homologous chromosomes are represented in this lane. If the two homologous chromosomes have similar DNA sequences at this site, then the resulting fragments will be of equal length and therefore will generate the pattern of one single dark band as for either person 2 or person 3; whereas if the DNA sequences at

both homologous chromosomes are different, then the resulting fragments will be of different length which will therefore generate the pattern of two dark bands as for person 1. Also as the site at which RFLPs are detected is short enough to segregate like a single locus, person 1 in this figure could then be seen as being heterozygote at this site (or locus) whereas person 2 or 3 could be considered homozygote. The genotypes of the RFLPs are thus quite similar to those of traditional markers and in practice most of them will appear as codominant genetic markers. The advent of this technique had two major consequences; if these RFLPs encode a gene of interest (disease) then this gene is isolated and its locus is known extremely closely but if they do not which is usually the case they can be used as valuable markers showing linkage to neighbouring disease loci. Also since a potentially large number of this kind of DNA sequence markers can be obtained, there is hope that eventually the whole human genome may be so densely populated by RFLPs that it will become possible to determine the chromosomal location of every human gene via linkage to the RFLPs.

1.4 Linkage analysis and 2-loci situation

A-Testing for linkage:

Statistical tests, designed to detect linkage between two loci, form a major component of linkage analysis. Throughout history of the subject many tests had been invoked. The most influential one of them on today's practice was introduced by Morton(1955). He based his method on the theory of sequential analysis as well as using the lod score statistics, $z(\theta)$, introduced by Haldane and Smith(1947). If $P(r|\theta)$ is the probability of obtaining the data r when the true recombination fraction is θ then the lod score, $z(\theta)$, is defined as:

$$z(\theta)= \log_{10}\left\{ \frac{P(r|\theta)}{P(r|0.5)} \right\} \qquad (1.1)$$

Where the name "lod" is actually an acronym for the "logarithm of the odds ratio". If $z_i(\theta)$ for i=1,2,.. is the lod score for several independent families then the sequential test (Wald 1947) introduced, in the context of linkage, by Morton will proceed as follows. The null hypothesis of a free recombination fraction (i.e) H0:θ=0.5 is tested against the simple alternative of θ being equal to a chosen value θ_1 (i.e) H1:$\theta=\theta_1$:$\theta_1<0.5$. Morton suggested using one of the four values θ_1= 0.05, 0.1, 0.2 and 0.3. The total lod score

$$Z(\theta_1)= \Sigma z_i(\theta_1) \qquad (1.2)$$

is determined after each new family has been investigated. Actually the data $r_1, r_2, ..$ of these families will accumulate until some stopping criterion is met. The test employs two positive numbers A>1 and B<1 and continues until either:

$Z(\theta_1)\geq \log A$ in which case H0 is rejected (i.e) $\theta<0.5$,

or $Z(\theta_1)\leq \log B$ in which case H1 is rejected (i.e) $\theta\neq\theta_1$.

Otherwise if $\log B<Z(\theta_1)<\log A$ no conclusion is made and more families are sampled. The sample needed until the test terminates is thus a random variable which depends on the true unknown θ. A and B are determined such that type I error$\leq\alpha$ and type II error$\leq\beta$, where α and β are given small positive constants. Under the assumption of a negligible excess over the boundaries at the solution, (i.e) assuming that $Z(\theta_1)\simeq\log A$ or $Z(\theta_1)\simeq\log B$ at the solution, then using $A\simeq(1-\beta)/\alpha$ and $B=\beta/(1-\alpha)$ will ensure the above requirement about type I and type II errors. Morton recommended the critical values logA=3 and logB=-2 which correspond to α=0.001 and β=0.01. Such a stringent significance level is chosen in order to be compatible with the low prior probability of autosomal linkage which could roughly be seen as

being less then 1/22≈0.05 (Actually 1/22 is a rough approximation of the prior probability of the two genes being on the same autosomal chromosome).

Nowadays, the test of the hypothesis of free recombination is carried out against the composite alternative hypothesis of linkage H1:$\theta<0.5$ using the general likelihood ratio test but in terms of the lod score. This method is known as the lod score method and proceeds as follows. The maximum likelihood estimate, MLE, $\hat{\theta}$, which is defined as that value of θ maximizing $Z(\theta)$, is determined. Then the hypothesis of no linkage will be rejected if $Z(\hat{\theta})\geqslant 3$. The critical point of 3 recommended by Morton is thus still being used, although today's test is neither directed against a simple alternative nor carried out in a strictly sequential manner.

Chotai(1984) gave a thorough discussion about the lod score method when it is either considered as a sequential or a fixed sample size test. He emphasized the fact that by using $Z(\hat{\theta})$ instead of $Z(\theta_1)$ in a sequential test, then the above formulas for the boundaries A and B are no longer applicable. He also pointed out that in terms of Morton's sequential test, the assumption of a negligible excess over the boundaries at the end of the test is not justified by the current collection of the data which is carried out in terms of groups of pedigrees. Seen as a fixed sample size test, he investigated the adequacy of the x^2 approximation of the generalised likelihood ratio test for some genetical data. Actually he studied the approximation of the significance level $\alpha= P\{Z(\hat{\theta})>\log A|\theta=0.5\}$ by $\alpha= P\{x_1^2>c^2\}$, where $c^2=2\ln A$, for pedigree data consisting of n double backcross matings (see later section(1.5)-B) which amounts to a binomial distribution with sample size n, probability of success $\phi=f(\theta)$

:$0<\phi<0.5$ and r number of successes. But since the null hypothesis is tested in a one sided manner against $H1:\theta<0.5$, then α should be obtained by dividing the significance level of the approximate x_1^2 by 2 (i.e) $\alpha \simeq 0.5P\{x_1^2>c^2\}$ _a fact pointed out by Ott(1977). Chotai calculated the exact significance level when α=0.001, 0.01, 0.05 and $n\leqslant 50$ for both the one sided and two sided approximation. He found out that the two sided test gave an adequate approximation even for small n ($n\geqslant 10$) and was safer to use than the one sided test. Actually at n=50 the exact type I error were 0.0013 and 0.0595 for the one sided test and roughly (read from a plot) equal to 0.0005 and 0.035 for the two sided test when α=0.001 and 0.05 respectively. Also, in order to remedy the fact of the sequential collection of the data in terms of groups of pedigrees, Chotai promoted the idea of applying the group sequential approach, used recently in clinical trials and which is based on the repeated sequential test of Armitage(1975), to linkage analysis.

B-Interval estimation:

A number of methods for constructing confidence intervals, under different assumptions, have been adopted and used in linkage analysis. Morton(1956) based his method on the approximation of the observed likelihood $\Pi_i P(r_i|\theta)$ by a normal density with mean $\theta_o=\hat{\theta}$ and a variance σ^2given that n is large enough. Given this assumption, $Z(\theta)=\Sigma z_i(\theta)$ would be approximately quadratic in θ:

$Z(\theta)= a+b\theta+c\theta^2$ where $b=M\hat{\theta}/\sigma^2$ $c=-M/2\sigma^2$ $M=\log_{10}e$

Any three points say $(\theta_1,z_1),(\theta_2,z_2)$ and (θ_3,z_3) would be sufficient to determine a,b,c and therefore $\hat{\theta}$ and σ^2. Then, a large sample confidence interval for θ would be given by:

$\hat{\theta}-t\sigma<\theta<\hat{\theta}+t\sigma$, where t is a standard normal percentile with

confidence coefficient $(1-\alpha)$. Ott(1985) reintroduced this method and generalised it for more than one parameter.

Another kind of interval estimate based on the asymptotic distribution of $2\ln P(r|\hat{\theta})/P(r|\theta_o)$ which is χ^2_1 when $\theta=\theta_o$ is defined as $\theta_1<\theta_o<\theta_2$ where θ_1 and θ_2 are the intersection of the horizontal line $Z= Z(\hat{\theta})-3$ (or sometimes $Z=Z(\hat{\theta})-2$) with the curve of the lod score $Z(\theta)$ _Ott(1977,1985).

The final kind of interval estimate that is going to be mentioned here is adopted from Morton's sequential test. It is defined as $\theta_1<\theta\leqslant 0.5$ where θ_1 is that value of θ with a lod score of (-2) (i.e) $Z(\theta_1)=-2$. In the context of a fixed sample size and due to the fact that this interval never allows us to exclude values of θ lying between $\hat{\theta}$ and 0.5, Ott(1985) suggested to only use it when the test of linkage is not significant. He also explained that this interval is chosen to be conservative due to the low prior probability of linkage.

Some of these intervals plus others will be discussed later in details in the fourth and fifth chapter of this thesis.

1.5 Likelihood for pedigree data and some mating types

A-Likelihood for pedigree data

Writing the likelihood as an explicit function of θ is not an easy task when dealing with large family pedigree data. For a family of size n, let y_i and g_i be the phenotype and genotype of the i^{th} member of the family. The likelihood function is defined as the probability with which the given phenotypes of the family can occur, it could be written as follows:

$$L(\theta)= P(y_1 y_2 .. y_n|\theta)$$

$$= \sum_{g_1}\ldots\sum_{g_n} P(y_1..y_n|g_1..g_n\ )P(g_1..g_n\ |\theta) \quad (1.3)$$

Where the multiple sum in (1.3) is over all possible genotypes for each individual in the family. In general the phenotypes are

not mutually independent but given the genotypes including phase they are conditionally independent (i.e) $P(y_1..y_n|g_1..g_n)=\Pi_i P(y_i|g_i)$. Also, the genotypes of all offsprings will be mutually independent only given their parents' genotypes. Therefore $P(g_1..g_n)$ could be written as $\Pi_i P(g_i|g_{m(i)}g_{f(i)})$ where $P(g_i|g_{m(i)}g_{f(i)})$ represents the probability of an individual genotype given his mother's and father's genotypes. If the parents of this individual are part of the pedigree data then this probability will be function of θ, otherwise $P(g_i|g_{m(i)}g_{f(i)})$ will be replaced by $p(g_i)$ which will be calculated from population gene frequencies. $L(\theta)$ could therefore be written as :

$$L(\theta)= \sum_{g_1}\cdots\sum_{g_n} \prod_i P(y_i|g_i)P(g_i|g_{m(i)}g_{f(i)}) \qquad (1.4)$$

$P(y_i|g_i)$ is the probability of occurence of the i^{th} individual's phenotype given his genotype, it is mainly taken to be either 0, if the phenotype is incompatible with the genotype, or 1 if it is. Thus the multiple sum in (1.4) will be reduced to just the sum of the probabilities of all genotypes compatible with the phenotypes. Nevertheless, the evaluation of the likelihood as it is written will be time consuming even when computers are used. Elston and Stuart(1971) proposed a highly efficient algorithm which will calculate the likelihood in a recursive manner. Later computer programs were written, making use of this algorithm, to calculate the likelihood. The program LIPED written by Ott(1974) and LINKAGE written by Lathrop and Lalouel(1984), are frequently used when dealing with family pedigrees involving two loci only.

An example from the genetical literature is used here to illustrate the derivation of the likelihood using formula (1.4). Figure(1.7) shows the pedigree data of a family, named family R after Morton (1956), suffering from a rare autosomal dominant

hereditary disease known as Elliptocytosis ,which is suspected to be linked to the Rhesus, Rh, blood locus. In this pedigree the only Rh genes which appear to be segregating are R_1 and r, the Rh phenotype of each individual is shown in the figure below his or her symbol. On the other hand due to the rarity of the disease, the affected individuals which are represented by dark symbols are assumed to be heterozygote at the disease locus, (i.e) their genotype is assumed to be El/el, where El represents the dominant disease allele and el its corresponding normal one. In this pedigree n=19 and the numbering of each individual is shown in the figure below his or her symbol. The phenotype and possible genotypes of each individual at both the Rh and disease loci are shown in table(1.3). The genotypes are chosen so that they are compatible with both the phenotype of the individual, (i.e) such that $P(y_i|g_i)=1$, and with the genotypes of his parent, (i.e) such that $P(g_i|g_{m(i)}g_{f(i)})\neq 0$. From the table we can see that only individuals 1,7,15 and 19 have more than one possible genotype. Therefore the summation in (1.4) will be reduced from n=19 to just four summations. The likelihood of this pedigree could therefore be written as follows:

$$L(\theta)= \sum_{g_1}\sum_{g_7}\sum_{g_{15}}\sum_{g_{19}}\prod_{1}^{19} P(g_i|g_{m(i)}g_{f(i)})$$

$$= P(g_2)P(g_8)P(g_{16})$$

$$\times \sum_{g_1}\Big\{ \Big[\prod_{i=3}^{6} P(g_i|g_1,g_2)P(g_{13}|g_1,g_2)P(g_{14}|g_1,g_2)\Big]$$

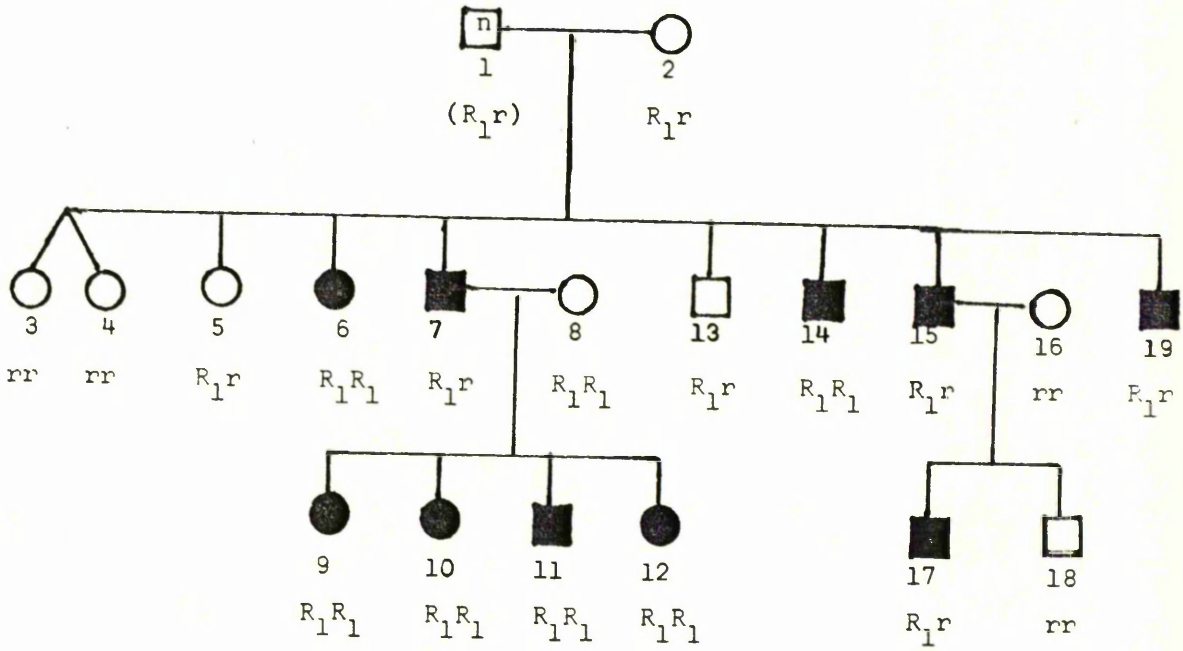
$$\times \Big[\sum_{g_7} P(g_7|g_1,g_2)\prod_{i=9}^{12} P(g_i|g_7,g_8)\Big]$$

$$\times \Big[\sum_{g_{15}} P(g_{15}|g_1,g_2)\prod_{i=17}^{18} P(g_i|g_{15},g_{16})\Big]$$

$$\times \Big[\sum_{g_{19}} P(g_{19}|g_1,g_2)\Big]\Big\}$$

Let the genotype R_1El/rel and R_1el/rEl be denoted by genotype 1

Figure(1.7) Pedigree R, which is affected with the rare autosomal dominant disease of Elliptocytosis



(Note that the phenotype of individual 1 is not known but deduced from his progeny).

Table (1.3) The phenotypes and possible genotypes of the individuals in figure(1.7)

i^{th} Individual	Phenotype	Possible genotypes
1	R_1r-Elel	R_1El/rel or R_1el/rEl
2	R_1r-elel	R_1el/rel
3	rr -elel	rel/rel
4	rr -elel	rel/rel
5	R_1r-elel	R_1el/rel
6	R_1R_1-Elel	R_1El/R_1el
7	R_1r-Elel	R_1El/rel or R_1el/rEl
8	R_1R_1-elel	R_1el/R_1el
9	R_1R_1-Elel	R_1El/R_1el
10	R_1R_1-Elel	R_1El/R_1el
11	R_1R_1-Elel	R_1El/R_1el
12	R_1R_1-Elel	R_1El/R_1el
13	R_1r-elel	R_1el/rel
14	R_1R_1-Elel	R_1El/rel
15	R_1r-Elel	R_1El/rel or R_1el/rEl
16	rr -elel	rel/rel
17	R_1r-Elel	R_1El/rel
18	rr -elel	rel/rel
19	R_1r-Elel	R_1El/rel or R_1el/rEl

and genotype 2 respectively and also denote the 1st, 2nd, 3rd and 4th parentheses in the above formula by I_1, I_2, I_3 and I_4 respectively. Now if $g_1=1$, then $P(g_3|g_1,g_2)$ (where from table(1.3) g_3 is rel/rel) will be equal to the probability that individual 3 had received haplotype rel from her mother with probability 0.5 and had received haplotype rel from her affected father with probability $(1-\theta)/2$. By carrying on like that for each individual, I_1, I_2, I_3 and I_4 could be found to be:

<u>for $g_1=1$</u> $\quad I_1 = (1-\theta)^4/2^{12}$, $I_2=[\theta^5+(1-\theta)^5]/2^6$

$\quad I_3 = [\theta^3+(1-\theta)^3]/2^4$, $I_4=1/4$

<u>for $g_1=2$</u> $\quad I_1 = \theta^4/2^{12}$, $I_2=\theta(1-\theta)[\theta^3+(1-\theta)^3]/2^6$

$\quad I_3 = \theta(1-\theta)/2^4$, $I_4=1/4$

So that $L(\theta)$ would be:

$$L(\theta)=\text{constant} \times \{(1-\theta)^{12}+\theta^3(1-\theta)^9+\theta^5(1-\theta)^7+\theta^6(1-\theta)^5+\theta^8(1-\theta)^4+\theta^9(1-\theta)^2\}$$

B-<u>Likelihood and mating types</u>:

For some other kind of pedigrees, calculating the likelihood will be a straightforward procedure. For a mating to be informative for linkage, when two loci are involved, at least one of the two parents has to be doubly heterozygote. Depending on the genotype of the other parent, different mating types could be distinguished. If he (or she) is doubly homozygote, singly heterozygote or doubly heterozygote, then the mating type will be termed a *double backcross*, a *single backcross* or a *double intercross*. But to calculate the likelihood we will still have to distinguish between cases where the phase of a doubly heterozygote parent is known or not.

The following general scheme could be followed to calculate the likelihood for such mating, <u>when the phase is known</u>. (Note that the phase will only be known for certainty in some three

generation family pedigree data, where the phase of the parents can be deduced from the grand parents _as in a previous example shown in figure(1.5)). Firstly, a list of all possible offsprings genotypes, that this mating can produce, with their corresponding probability of occurence will be prepared. Secondly, depending on the mode of inheritance (dominant,codominant,..etc), the different phenotypes and their probabilities will be written. Usually different phenotypes will have the same probability of occurence, so the final step will be to combine all these phenotypes into one class, ending up with different phenotype classes with different probabilities, say $p_1, p_2..p_k$. If such mating produces n offsprings with $r_1, r_2, ..r_k$ offsprings in each of the k classes, then their likelihood function will be proportional to

$$p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k}$$

(Note that the n offspring's phenotypes are mutually independent because the genotypes of their parents are fully known).

Applying the above scheme to a phase known single backcross mating AB/ab×Ab/ab is given here as an example. In table(1.4), which is a two way table whose first row and first column show the four possible haplotypes produced by the doubly heterozygote parent and the two possibe haplotypes of the singly heterozygote parent with their corresponding probabilities respectively, a list of all possible genotypes are presented. Under codominance mode of inheritance, these genotypes will correspond to six different phenotypes which are shown in table(1.5) with their corresponding probabilities. Out of the six phenotypes, only three different classes of phenotypes can be distinguished, they are shown in table(1.6). If this mating or any other single backcross mating produces r_1, r_2, r_3 offsprings with phenotype

PHASE KNOWN

Mating AB/ab × Ab/ab, codominant inheritance

Table (1.4) Offsprings' genotypes

		AB	Ab	aB	ab
		(1-θ)/2	θ/2	θ/2	(1-θ)/2
Ab	0.5	AB/Ab	Ab/Ab	aB/Ab	ab/Ab
ab	0.5	AB/ab	Ab/ab	aB/ab	ab/ab

Table (1.5) Offsprings' phenotypes

i^{th}	phenotype	Probability
1	AA-Bb	(1-θ)/4
2	AA-bb	θ/4
3	Aa-Bb	1/4
4	Aa-bb	1/4
5	aa-Bb	θ/4
6	aa-bb	(1-θ)/4

Table (1.6) Offsprings' phenotype classes

Class k	i or j	Probability p_k
1	(1 or 6)	(1-θ)/2
2	(2 or 5)	θ/2
3	(3 or 4)	1/2

(Note that i or j in table(1.6) refer to the numbering in table(1.5) of the phenotypes).

class 1,2 and 3 respectively then the probability distribution of the data could be seen as follows

$$r_1\ r_2\ r_3 \sim \text{Multinomial}(n;\ (1-\theta)/2,\ \theta/2,\ 0.5)$$

By analogy the probability distribution of the other matings can be deduced. Table(1.7) gives the probability distribution of n offsprings produced by different mating type when the phase is known and codominance mode of inheritance is assumed. For a more elaborate discussion of this subject and for other type of mating and different mode of inheritance the reader is referred to Ott(1985) section (3.6).

When dealing with an unknown phase mating, we have to take all possible phases into account. For a double backcross mating, the doubly heterozygote parent could have one of the two equally likely phases, phase 1= AB/ab or phase 2= Ab/aB. Under this mating four possible genotypes, g_i i=1,2,3,4 (where the subscript i is used here to index the type of genotype and not the individual), will be produced. Table(1.8) shows the conditional probabilities of them given each of the two phases. Under either codominance or dominance (with A and B denoting the dominant alleles) mode of inheritance the four genotypes will correspond to four phenotypes. If such mating produces only one offspring, then the probability of his (or her) phenotype will be uninformative for θ, where

$$P(x_i) = \Sigma_j\ P(x_i|\text{phase } j)P(\text{phase } j)$$
$$= 0.25$$

If two offsprings are produced the situation will change. The total number of possible phenotypes for the two offsprings will be equal to 4+(4)(3)/2=10. For each possibility we can calculate the corresponding probability of occurence. For example if both children have phenotype Aa_Bb then the probability of occurence

Table (1.7) Probability distribution of n offsprings from phase known matings

Mating		Probability distibution
Double backcross	AB/ab×ab/ab	$r \sim bi(n;\theta)$
Single backcross	AB/ab×Ab/ab	$r_1 r_2 r_3 \sim Mult(n;\ 0.5\ (1-\theta)/2\ \theta/2)$
Double intercross	AB/ab×AB/ab	$r_1 r_2 r_3 r_4 \sim Mult(n;\ (1-\theta)^2/2\ 2\theta(1-\theta)\ \theta^2/2\ (\theta^2+(1-\theta)^2)/2)$

PHASE UNKNOWN

Double backcross mating, codominance inheritance

Table (1.8) Conditional probabilities of the offsprings' genotypes given the phase

i^{th} genotype	Phase 1	Phase 2
1 AB/ab	$(1-\theta)/2$	$\theta/2$
2 Ab/ab	$\theta/2$	$(1-\theta)/2$
3 aB/ab	$\theta/2$	$(1-\theta)/2$
4 ab/ab	$(1-\theta)/2$	$\theta/2$

Table (1.9) Offsprings' phenotype classes for families with two offsprings

class k	ij	Probability p_k
1	11,22,33,44,14,23	$\theta^2+(1-\theta)^2$
2	12,13,24,34	$2\theta(1-\theta)$

(Note that ij in table(1.9) refer to the numbering in table(1.8) of the genotypes interpreted here as phenotypes)

will be $((1-\theta)/2)^2(0.5)+(\theta/2)^2(0.5)=[\theta^2+(1-\theta)^2]/8$. By evaluating the probability of all 10 possibilities it turns out that only two different probabilities occur and thus producing only two phenotype classes. Table(1.9) shows these two phenotype classes,their corresponding probabilities along with the possible combination of the offspring's phenotypes which can produce them. If we have n such matings with each one producing two offsprings, known as sibpair of offsprings, then if r is the number of sibpairs with phenotype class 2. then the probability distribution of the data could be seen as follows

$$r \sim bi(n;\ \phi) \quad \text{where } \phi = 2\theta(1-\theta) \qquad (1.5).$$

1.6 Three loci or more

1.6.1 General background

A-Introduction:

Traditionally, linkage analysis for a disease locus versus n marker loci were carried out as a sequence of two point analyses, which have been explained in a section(1.4). For each comparison between a disease locus and i^{th} marker locus, likelihood and lod score are calculated and treated as being independent of those for the other comparisons. But with the linkage map growing denser, simultaneous analysis of three loci (or more) becomes more important.

In genetic linkage with two loci situation a single parameter, the recombination fraction between them θ, is of interest. With three loci, (eg) A,B and C, three recombination fractions , θ_{ab}, θ_{bc} and θ_{ac} are of interest; they are known as the marginal recombination fractions and denote the probability of a recombination occuring in segment AB, segment BC and segment AC respectively. However, given an informative parental genotype at the three loci, (i.e) a triple heterozygote parent, the

haplotypes which could be produced by this parent (and therefore the observed offsprings), can be seen as events of four types of multiple recombination, type I which shows a recombination in segment AB and segment BC, type II which shows a recombination in segment AB and no recombination in segment BC, type III which shows no recombination in segment AB and a recombination in segment BC and finally type IV which shows no recombination in both segments. As an example, if the parent's genotype is ABC/abc, with the loci arranged in that order, then a haplotype of the form (AbC or aBc) and (aBC or Abc) and (ABc or abC) and (ABC or abc) will correspond to the 1^{st}, 2^{nd}, 3^{rd} and 4^{th} type of multiple recombination respectively.

In general, let the probability of these four types be denoted by α, β, γ and δ respectively. Since $\alpha+\beta+\gamma+\delta=1$, there are three independent parameters, say α, β and γ, from which the marginal recombination fractions could be calculated. As θ_{ab} is the probability of a recombination in AB whatever happens in the other segment then $\theta_{ab}=\alpha+\beta$ and by analogy $\theta_{bc}=\alpha+\gamma$. On the other hand a recombination in AC which corresponds to an odd number of crossing-over between A and C, corresponds to either a recombination in AB and no recombination in BC, with probability β, or to a recombination in BC and no recombination in AB, with probability γ, therfore $\theta_{ac}=\beta+\gamma$. In summary, the likelihood function of a three loci situation can be expressed in term of either α, β and γ or θ_{ab}, θ_{bc} and θ_{ac} where

$$\begin{array}{lcl} \theta_{ab}=\alpha+\beta & & 2\alpha=\theta_{ab}+\theta_{bc}-\theta_{ac} \\ \theta_{bc}=\alpha+\gamma & \leftrightarrow & 2\beta=\theta_{ab}-\theta_{bc}+\theta_{ac} \\ \theta_{ac}=\beta+\gamma & & 2\gamma=\theta_{ac}+\theta_{bc}-\theta_{ab} \end{array} \qquad (1.6)$$

But since the probability of the multiple recombination events α, β and γ must be non negative, then the marginal recombination

fractions must satisfy the triangle inequality, (i.e) the sum of any two of them must be greater or equal to the third. Also, any marginal recombination must be restricted to the range of [0.0,0.5]. Therefore the likelihood function will be restricted to the following range of the parameters

$$\begin{array}{lcl} 0<\alpha+\beta<0.5 & & 0<\theta_{ac}<\text{Min}[0.5;\theta_{ab}+\theta_{bc}] \\ 0<\alpha+\gamma<0.5 & \leftrightarrow & 0<\theta_{bc}<\text{Min}[0.5;\theta_{ab}+\theta_{ac}] \\ 0<\beta+\gamma<0.5 & & 0<\theta_{ab}<\text{Min}[0.5;\theta_{ac}+\theta_{bc}] \end{array} \quad (1.7)$$

Note that, relation(1.6) and restriction(1.7) are true for any order of the three loci. But for a given order, (eg) ABC, one would require at least one additional restriction, namely that θ_{ac}, the recombination fraction between the flanking loci, be at least as large as either θ_{ab} or θ_{bc} (Ott 1985, page171). Therefore, in terms of the θs, restriction(1.7) for the order ABC will be:

$$\text{Max}[\theta_{ab};\theta_{bc}]<\theta_{ac}<\text{Min}[0.5;\theta ab+\theta_{bc}] \quad (1.8)$$

B-Interference:

Another phenomenon, that we have to take into account when we are considering linkage analysis of three loci (or more), is the phenomenon of interference. In genetics it is well established that the pattern of crossing-over, and therefore recombination fraction, in any segment of a chromosome is not independent of the pattern of crossing-over in any other segment. The failure of random crossing-over along the chromosome is the phenomenon of *interference*. Generally, the occurence of a point of exchange tends to inhibit the formation of other such points in its neighbourhood, so called positive interference. A convenient index of the strength of interference is measured by the coefficient of coincidence, c, which is defined as follows

$$c= \frac{\alpha}{(\alpha+\beta)(\alpha+\gamma)} = \frac{\theta_{ab}+\theta_{bc}-\theta_{ac}}{2\theta_{ab}\theta_{bc}} , \qquad (1.9)$$

given the gene order ABC. If there is no interference and crossing-over occurs at random, then the probability of a recombination in AB and in BC, α, will be equal to the product of the marginal recombination θ_{ab} and θ_{bc}, (i.e) $\alpha=(\alpha+\beta)(\alpha+\gamma)$ and c=1. With positive interference, however, the frequency of double recombination will be less than the random value, so that $\alpha<(\alpha+\beta)(\alpha+\gamma)$ and c<1. From (1.9), θ_{ac} could be written in terms of θ_{ab}, θ_{bc} and c as follows

$$\theta_{ac}= \theta_{ab}+\theta_{bc}-2c\theta_{ab}\theta_{bc} \qquad (1.10)$$

Therefore, given positive interference, (i.e) $c\leqslant 1$, the lower bound of θ_{ac} in the restriction (1.8) will be sharpen as follows

$$\theta_{ab}+\theta_{bc}-2\theta_{ab}\theta_{bc}<\theta_{ac}<Min[0.5;\theta_{ab}+\theta_{bc}] \qquad (1.11)$$

We will still have to restrict, θ_{ab} and θ_{bc} to the range of [0.0;0.5]. Also, if we want to write down the likelihood in terms of θ_{ab},θ_{bc} and c, when the gene order is ABC and positive interference is assumed then the multiple recombination probabilities could be written as follows

$$\begin{aligned} \alpha &= c\theta_{ab}\theta_{bc} \\ \beta &= \theta_{ab}(1-c\theta_{bc}) \\ \gamma &= \theta_{bc}(1-c\theta_{ab}) \end{aligned} \qquad (1.12)$$

with restriction

$$\begin{aligned} &0<\theta_{ab}<0.5 \\ &0<\theta_{bc}<0.5 \\ &Max\left[0;\ \frac{\theta_{ab}+\theta_{bc}-0.5}{2\theta_{ab}\theta_{bc}}\right]<c\leqslant 1 \end{aligned} \qquad (1.13)$$

C-Mating types:

Under three loci situation, two kinds of mating are going to be used in this study, the phase known triple backcross mating

ABC/abc×abc/abc, which we will call here, for convenience, mating I and the phase unknown triple backcross mating called here mating II. In analogy to the two loci situation, we produced table(1.10) and table(1.11) to explain the possible outcomes of mating I and table(1.12) and table(1.13) to explain mating II. In table(1.10), we can see the eight possible haplotypes which could be produced by the triple heterozygote parent and the one possible haplotype of the homozygote parent along with their corresponding probabilities. Therefore only eight types of genotype could be produced by this mating. By assuming either codominance or dominance mode of inheritance (with A, B and c denoting the dominant alleles) at each of the three loci, the eight genotype symbols may be interpreted as symbols of possible phenotypes. Combining into one class the different phenotypes which have equal probabilities leads to table(1.11), comprising four phenotype classes with different probabilities. As seen before, if such mating produces n offsprings with r_1, r_2, r_3 and r_4 denoting the number of offsprings with phenotype class 1,2,3 and 4 respectively, then the probability distribution of such data will have the following multinomial distribution:

$$r_1\ r_2\ r_3\ r_4 \sim \text{Mult}(n;\ \alpha,\beta,\gamma,\delta) \qquad (1.14)$$

For mating II, the triple homozygote parent will be abc/abc, the triple heterozygote parent on the other hand will have one of the following four equally likely phases: phase 1 ABC/abc, phase 2 ABc/abC, phase 3 AbC/aBc and phase 4 Abc/aBC. As in mating I, eight possible genotypes could be produced, which will also correspond to eight possible phenotypes under either dominant or codominant mode of inheritance. Shown in table(1.12) the conditional probabilities of these eight genotypes (phenotypes) given each of the four phases of the heterozygote parent. If such

PHASE KNOWN-3 LOCI SITUATION

Mating ABC/abc × abc/abc, codominance inheritance

Table (1.10) Offsprings' genotypes

		abc 1	genotype number
ABC	$\delta/2$	ABC/abc	1
ABc	$\gamma/2$	ABc/abc	2
aBC	$\beta/2$	aBC/abc	3
AbC	$\alpha/2$	AbC/abc	4
aBc	$\alpha/2$	aBc/abc	5
Abc	$\beta/2$	Abc/abc	6
abC	$\gamma/2$	abC/abc	7
abc	$\delta/2$	abc/abc	8

Table (1.11) Offsprings' phenotype classes

Class k	i or j	Probability p_k
1	4 or 5	α
2	3 or 6	β
3	2 or 7	γ
4	1 or 8	δ

(note that i or j in table(1.11) refer to the numbering in table(1.10) of the genotypes interpreted here as phenotypes).

PHASE UNKNOWN- 3 LOCI SITUATION

Triple backcross mating, codominance inheritance

Table (1.12) Conditional probabilities of the offsprings' genotypes given the phase

i^{th} genotype	Phase 1	Phase 2	Phase 3	Phase 4
1 ABC/abc	$\delta/2$	$\gamma/2$	$\alpha/2$	$\beta/2$
2 ABc/abc	$\gamma/2$	$\delta/2$	$\beta/2$	$\alpha/2$
3 aBC/abc	$\beta/2$	$\alpha/2$	$\gamma/2$	$\delta/2$
4 AbC/abc	$\alpha/2$	$\beta/2$	$\delta/2$	$\gamma/2$
5 aBc/abc	$\alpha/2$	$\beta/2$	$\delta/2$	$\gamma/2$
6 Abc/abc	$\beta/2$	$\alpha/2$	$\gamma/2$	$\delta/2$
7 abC/abc	$\gamma/2$	$\delta/2$	$\beta/2$	$\alpha/2$
8 abc/abc	$\delta/2$	$\gamma/2$	$\alpha/2$	$\beta/2$

Table (1.13) Offsprings' phenotype classes for families with two offsprings

Class k	i,j	p_k
1	11,22,33,44,55,66,77,88,45,36,27,18	$\alpha^2+\beta^2+\gamma^2+\delta^2$
2	12,34,56,78,17,28,35,46	$2(\alpha\beta+\gamma\delta)$
3	13,16,24,25,38,47,57,68	$2(\alpha\gamma+\beta\delta)$
4	14,15,23,26,37,48,58,67	$2(\alpha\delta+\beta\gamma)$

(Note that i and j in table(1.13) refer to the numbering in table(1.12) of the genotypes interpreted here as phenotypes).

mating produces only one offspring, then the unconditional probability of his or her phenotype will be (1/8), (i.e) uninformative for α,β or γ. Instead if two offsprings are produced, the total number of their possible phenotypes is equal to 8+(8)(7)/2=36, each of which will be informative about the parameters. For example if the first child has phenotype Aa-Bb-Cc and the second has phenotype aa-Bb-Cc then the corresponding probability of occurence of these phenotypes will be

$= (1/4)(\delta/2)(\beta/2)+(1/4)(\gamma/2)(\alpha/2)+(1/4)(\gamma/2)(\alpha/2)+(1/4)(\delta/2)(\beta/2)$

$= (\delta\beta+\alpha\gamma)/8$

Disregarding the order of the two offsprings, the probability of all possible 36 phenotypes can be evaluated. It then turns out that only four different probabilities occur. Table(1.13) shows these four phenotype classes, their corresponding probabilities along with the possible combination of offspring's phenotypes which can produce them. Again if we have n such matings with two offsprings each then, if r_1,r_2,r_3 and r_4 are the number of sibpairs with phenotype class 1, 2, 3 and 4 respectively then the probability distibution of such data will be

$$r_1\ r_2\ r_3\ r_4 \sim \text{Mult}(n;\ p_1,p_2,p_3,p_4) \qquad (1.15)$$

where the ps are as determined in table(1.13).

D-Map distance and map function

So far, the only measure of distance between any two loci was based on the recombination fraction between them. Actually this measure lacks the important additive property of any measure of distance in its stricter sense. Actually, if we have three loci A, B and C, arranged in that order then from (1.10) θ_{ac} will only be equal to the sum of the other two recombination fractions if c=0. This case corresponds to complete interference where a point of exchange completely inhibits the formation of other points in

its neighbourhood, and is only assumed to be true for groups of loci that are fairly near each other.

In genetics, a better scale of measurement between any two loci, known as the *map distance*, x, is defined as the average number of crossing-over occuring between the two loci.Enjoying the same properties of an average, this quantity will be automatically additive even if interference occurs, (i.e) $x_{ac}=x_{ab}+x_{bc}$ for order ABC whatever the value of c. The main disadvantage of this measure is that it can not be directly observed and must be deduced from the recombination fraction on the basis of suitable assumption. An important part of the study of genetic mapping has been dedicated to finding the relation between the recombination fraction and the map distance or what is usually called the *map function* f(.) where $x=f(\theta)$.

Two ways have been established in approaching this problem. The first method is to construct a mathematical model of the process of crossing-over and then compare its prediction with observations, most of which are collected from experimental genetics on the Drosophila's chromosome during reproduction. The best known formula was given by Haldane(1919) and was based on the assumption of crossing-over occuring randomly and independently along the chromosome, which actually condradicts empirical evidence. Nevertheless, Haldane's formula constitutes an important cornerstone in genetic mapping, as it is easy to apply to any number of loci studied along a certain chromosome (see later, subsection E page 41). Under the above assumption, the number of points of exchange occuring between two loci A and B at meiosis will have a Poisson distribution with parameter x. The probability of exactly r points of exchange occuring is accordingly:

$$P(r|x)= \frac{x^r e^{-x}}{r!} \qquad r=0,1,2...$$

The recombination fraction, θ, being the probability of an odd number of exchanges can easily be derived as follows

$$\theta= \sum_{r=o}^{\infty} P(2r+1|x)= \frac{1}{2}\left[\sum_{r=o}^{\infty} \frac{x^r e^{-x}}{r!} - \sum_{r=o}^{\infty} \frac{(-x)^r e^{-x}}{r!}\right]$$

$$\leftrightarrow \qquad \theta= 0.5(1-e^{-2x}) \qquad (1.16)$$

$$\leftrightarrow \qquad x= -0.5\ln(1-2\theta) \qquad (1.17)$$

(Note that under this map function, c=1)

Also, due to Haldane(1919), a basic differential equation relating the map distance and recombination fraction was derived. He assumed that any recombination fraction θ can be regarded as a function $\theta(x)$ of the map distance x of the segment in question, independent of the actual siting of the segment. For three loci A, B and C arranged in that order, and from (1.10), θ_{ac} is

$$\theta_{ac}= \theta_{ab}+\theta_{bc}-2c\theta_{ab}\theta_{bc}$$

Now, let

$$\theta_{ab}= \theta(x)$$

$$\theta_{bc}= d\theta$$

$$\theta_{ac}= \theta(x+dx)$$

then it follows that

$$\theta(x+dx)= \theta(x)+d\theta-2c_m(\theta)\theta(x)d\theta \qquad (1.18)$$

where $c_m(\theta)$ is Haldane's marginal coincidence relating to a finite interval with a very short adjacent interval. Since for very short interval θ and x are approximately equal, then $d\theta$ in (1.18) may be replaced by dx, then it follows that

$$\frac{\theta(x+dx)-\theta(x)}{dx} = 1-2c_m(\theta)\theta(x)$$

proceeding to the limits gives

$$\frac{d\theta}{dx} = 1-2c_m(\theta)\theta \qquad \leftrightarrow \qquad \frac{dx}{d\theta} = \frac{1}{1-2c_m(\theta)\theta} \qquad (1.19)$$

therefore, $$x(\theta)=\int_0^{\theta}\frac{du}{1-2c_m(u)u}$$

(Note that this reduces to (1.17) if $c_m(u)=1$).

The second method for finding $x(\theta)$ is based on finding a formula which fits empirical results. Biological results suggest a function that gives complete interference at very short distances and no interference at large distances, (i.e)

$c_m \to 0$ when $\theta \to 0$

and $c_m \to 1$ when $\theta \to 0.5$

The simplest function satisfying these conditions which is given by $c_m=2\theta$, was suggested by Kosambi(1944) the founder of this second approach. By using the Haldane's differential equation (1.19), the Kosambi map function can be easily found to be

$$x=\frac{1}{4}\ln\frac{1+2\theta}{1-2\theta} \quad \text{or} \quad x=\frac{1}{2}\tanh^{-1}(2\theta) \qquad (1.20)$$

Under this map function and by using formula(1.9) when θ_{ac} is replaced by

$$\theta_{ac} = f^{-1}(x_{ac}) = f^{-1}(f(\theta_{ab})+f(\theta_{bc})) \qquad (1.21),$$

c can be easily found to be

$$c=\frac{2(\theta_{ab}+\theta_{bc})}{1+4\theta_{ab}\theta_{bc}} \qquad (1.22)$$

Carter and Falconer(1951) suggested the choice of $c_m=(2\theta)^3$ which leads to

$$x=0.25(\tan^{-1}(2\theta)+\tanh^{-1}(2\theta)) \qquad (1.23)$$

Rao et al(1977) combined all the map functions mentioned so far in a general formula and then used data on meiosis in the human male to estimate a mapping parameter p. The suggested general formula is

$$x=(1/6)\{p(2p-1)(1-4p)\ln(1-2\theta)+16p(p-1)(2p-1)\tan^{-1}(2\theta)+$$
$$2p(1-p)(8p+2)\tanh^{-1}(2\theta)+6(1-p)(1-2p)(1-4p)\theta\} \qquad (1.24)$$

They estimated p by means of non linear least square methods to be equal to 0.35; a value intermediate between the Kosambi and Carter and Falconer map functions which correspond to p=0.5 and p=0.25 respectively, whereas when p=1 or p=0 formula (1.20) will correspond to no interference (Haldane) or complete interference respectively.

In order to compare the performance of the different map functions on multipoint data, Pascoe and Morton(1987) fitted all the above map functions plus others, not mentioned here, to two sets of data on the Drosophila X chromosome, the first set of data involved seven loci whereas the second involved nine loci. Two map functions fitted the data best, the Rao et al with p=0.33 and a new map function suggested by the authors themselves and called equation (3), it was based on the choice of $c_m=(2\theta)^2$ which led to

$$x=\frac{-1}{12}\ln\frac{(1-2\theta)^2}{(1+2\theta+4\theta^2)}+\frac{\sqrt{3}}{6}\tan^{-1}\frac{(1+4\theta)}{\sqrt{3}}-0.15115 \qquad (1.25)$$

Also in their paper, Morton and Pascoe generalised and used what is called the interval Markov assumption in order to relate the various multiple recombination events produced by data from seven loci and nine loci to the marginal recombination fractions (see next section).

Given a certain map function, the likelihood for three point data will be function of two parameters only ,θ_{ab} and θ_{bc}; this is due to the fact that given a certain map function, c or θ_{ac} will be function of the other two marginal recombination fractions. Ott(1985) questioned the practicality of estimating c from human three points data against assuming a plausible map function. He calculated the asymptotic standard error of the parameter estimates $\hat{\theta}_{ab}$ $\hat{\theta}_{bc}$ and $\hat{c}$ and the correlation among them

for phase known triple backcross families with one offspring each and phase unknown triple backcross with two offsprings each. The standard error and correlation are function of the true parameters, so under different combination of θ_{ab}, θ_{bc} and c at the Kosambi level, he found that the standard error of $\hat{c}$ is much higher than that of $\hat{\theta}_{ab}$ and $\hat{\theta}_{bc}$ although it decreases as the values of θ_{ab}, θ_{bc} and c increases. For the same combinations of θ_{ab}, θ_{bc} and c, he also calculated the number of families, n, needed to detect a true $c<1$, when a likelihood ratio test testing $H0:c=1$ against $H1:c<1$ is carried out at a significance level $\alpha=0.05$ and when a power $(1-\beta)=0.80$ is aimed to be achieved. His results indicate that the test can be expected to be most powerful for moderate values of θ_{ab}, θ_{bc} and c, for which n reaches its minimum. But even so, more than 800 and more than 2000 phase known and phase unknown triple backcross mating are needed respectively. He concluded that assuming a plausible value of c rather than estimating it would be more practical in the light of the data available in human linkage.

In a similar manner, Lathrop et al(1985) investigated the bias obtained under the assumption of no interference in the estimate of the recombination fraction between the flanking loci ((i.e) θ_{ac} if the order is ABC), as well as the number of offsprings needed to obtain a mean square error for $\hat{\theta}_{ac}$, when c is to be estimated, equal to its corresponding one if c is assumed to be 1. Given that the true c is at the Kosambi level, they found that the bias of $\hat{\theta}_{ac}$ is always less than 10% of the true value. Also they found that the estimates obtained under the assumption of no interference have smaller mean square error than the unrestricted estimates for less than 500 offsprings in phase known and 1800 offsprings in phase unknown families, when true $\theta_{ab}=\theta_{bc}<0.1$ and

when the mating type is of the form A1B1C1/A2B2C2×A3B3C3/A4B4C4 if the phase is known (note that under this mating the number of alleles at each of the three loci is sufficiently large so that parents can be considered to carry four different alleles at a locus).

E-More than three loci:

The importance of the map distances and map function becomes apparent when we try to deal with more than three loci. For n loci, there is a total number of (n)(n-1)/2 marginal recombination fractions and the same number of map distances. Of these map distances n-1 are between adjacent loci, the remaining (n-1)(n-2)/2 are between any two non adjacent loci and may be inferred from the previous n-1 distances by using the additive property of this measure. When a suitable map function is chosen all map distances could be transformed into recombination fractions, thus reducing the number of parameters into n-1 independent ones. However, as seen with the three loci situation, the observed data are in terms of the multiple recombination events. In each of the n-1 adjacent segments a recombination may or may not occur, which makes the total number of these multiple events equal to 2^{n-1}, but since the sum of their probabilities is equal to one, the probabilities of these events are determined by $2^{n-1}-1$ independent parameters. The problem that we have to face now, is how to write down the $2^{n-1}-1$ multiple recombinations in terms of the (n)(n-1)/2 marginal recombinations which may then be reduced to n-1 map distances. For n=2 and n=3, $2^{n-1}-1=(n)(n-1)/2$ and no problem will arise. For n>3, $2^{n-1}-1>(n)(n-1)/2$ and additional assumptions have to be made. Before outlining some of the different approaches suggested so far to deal with this problem, let us use the following notation:

let θ_i and x_i be the recombination fraction and map distance in the i^{th} segment.

θ_{ij} be the recombination in the i^{th} and j^{th} segments; and similarly for θ_{ijk},...etc

θ_{i+j} be the recombination in the i^{th} or j^{th} segment, but not in both; and similarly for θ_{i+j+k},..etc.

As for the multiple recombination probabilities, let:

P_i be the recombination in the i^{th} segment only;

P_{ij} be the recombination in the i^{th} and j^{th} segments only; and so on for P_{ijk},P_{ijkl},...etc.

In general:

$$\left.\begin{aligned} \theta_i &= P_i + \sum_{j\neq i} P_{ij} + \sum_{k\neq i\neq j} P_{ijk} + \ldots \\ \theta_{ij} &= P_{ij} + \sum_{k\neq i\neq j} P_{ijk} + \ldots \\ &\ldots \text{ etc} \end{aligned}\right\} \qquad (1.26)$$

Over the years, several suggestions have been made to relate all the multiple events with the marginal recombinations, only some of which are going to be mentioned here:

(i)Assuming no interference ((i.e) using the Haldane map function). Under this assumption the probability of multiple recombination will be given by the product of the appropriate marginal recombination fractions

$$\text{(eg) } P_{125} = \theta_1\theta_2(1-\theta_3)(1-\theta_4)\theta_5\prod_{i=6}^{n-1}(1-\theta_i)$$

In this manner all the Ps will be automatically specified by the n-1 map distances.

(ii) Setting all probabilities of three or more simultaneous recombinations equal to zero. By doing so, the number of unknown Ps will be automatically reduced to (n)(n-1)/2, which could be easily related to the marginal recombinations. Afterwards any map

function could be used to express the recombination fractions in term of the n-1 map distances (Ott 1985).

(iii)Using the interval Markovian assumption. This method was introduced by Morton and MacLean(1984) and is based on the assumption that a crossover divides the chromosome into two segments between which there is no interference. Therefore, interference in a region is assumed to depend only on the nearest crossover. So that if we have three adjacent regions i,j and k in that order on the chromosome then the conditional probability of recombination in region k given recombination in region i and j is given by

$$P(k|ij) = P(k|j) \qquad (1.27)$$

The position of the crossover within j is ignored, (i.e) segment j is treated as a geometric point. By definition

$$P(k|ij) = \frac{\theta_{ijk}}{\theta_{ij}}$$

and $$P(k|j) = \frac{\theta_{jk}}{\theta_{j}}$$

So that the interval Markovian assumption implies that

$$\theta_{ijk} = \frac{\theta_{ij}\theta_{jk}}{\theta_{j}} \qquad (1.28)$$

This formula is exact if j is a geometric point ($x_j=0$) or if x_j is so great that $\theta_{ij}=\theta_i\theta_j$ and $\theta_{jk}=\theta_j\theta_k$ and therefore $\theta_{ijk}=\theta_i\theta_j\theta_k$. For intermediate values of x_j formula (1.28) will be at best an approximate one. By using this formula and further assuming that quadruple or more crossovers are negligible, Pascoe and Morton(1987) derived the necessary formulas relating the multiple recombination probabilities to the marginal ones.

(iv)Using the point Markovian assumption, this method was introduced by Bailey(1961 page157) and is based on the assumption that when a crossover point occurs it effectively divides the

chromosome into regions which do not interfere with each other although there may still be interference within each region. The difference between method(iii) and (iv) is that Bailey does not treat the segment within which the crossover occurs as a geometric point. By further assuming that we deal with several relatively short segments, Bailey derived in details the necessary formulas relating the multiple recombination probabilities with the marginal ones.

To give an idea about the implication of this assumption let us consider the four loci situation with order ABCD. By definition

$$\theta_{123} = P(\text{recombination in } 1^{st} \text{ \& in } 2^{nd} \text{ \& in } 3^{rd} \text{ segment})$$

Now if we consider a small increment δx situated at a point P interior to BC, then the recombination in BC will be divided into recombination in BP denoted by θ or a recombination in PC denoted by $(\theta_2-\theta)$, as a first order approximation. Also, note that $\theta_{123}=\int\delta\theta_{123}$, where:

$$\delta\theta_{123} = P(\text{recombination in } 1^{st} \text{ \& in } \delta x \text{ \& in } 3^{rd} \text{ segment})$$

If a crossover is established in δx then according to the point Markovian assumption

$$\delta\theta_{123} = P(\text{recomb. in } 1^{st} \text{ \& } \delta x)P(\text{recomb.in } 3^{rd} | \text{in } \delta x)$$

where both probabilities in the above expression could be written in terms of the marginal recombination $\theta_1,\theta_2,\theta_3,\theta$ and $\delta\theta$ (see Bailey page 157,158). Finally the multiple recombination θ_{123} could be then found by integrating $\delta\theta_{123}$ over θ when θ is varying between 0 and θ_2.

1.6.2 Linkage analysis

Linkage analysis for more than two loci consists mainly of two major investigations. Before discussing any of them, it is important to introduce another common measure used in linkage analysis, known as the *map location*, w_i, of locus i. This measure

determines the site of locus i on the genetic map and is defined as the map distance of this locus relative to a certain chosen point on the studied chromosome used as an orgin. Therefore if we have two loci A and B on the same chromosome then, $x_{ab}=|w_b-w_a|$ and if A is chosen as the origin, then $w_a=0.0$ and $w_b=\pm x_{ab}$. For any two loci data, the likelihood function could be written in term of this measure given a certain map function

(i.e) $L(\theta_{ab})= L(w_a,w_b)$ where $x_{ab}=f(\theta_{ab})=|w_b-w_a|$

Usually in genetic study the lod score $z(\theta_{ab})$ is of interest, $[z(\theta_{ab})=\log_{10}(L(\theta_{ab})/L(0.5))]$. The corresponding $z(w_b)$, where A is chosen as the origin, will be the log of the odds for w_b against a value large enough to imply no linkage. For multipoint data, on the other hand, the likelihood function, or the lod score, could only be written in terms of the map location parameters, if the overparameterisation problem, discussed in the previous section, is tackled by one way or another.

It is also worth mentioning that traditionally family pedigree data were collected for investigation about linkage between two loci, which along the years produced a big reserve of two loci data. If we want to use these data for multipoint analysis, then the overall likelihood will be the product of each of the two point likelihoods provided that the data on any two loci were collected from independent families. Therefore, if we have three loci A,B and C with two point likelihood $L_i(\theta_{ab})$, $L_j(\theta_{bc})$ and $L_k(\theta_{ac})$ on I,J and K independent families respectively, then the overall likelihood, $L(\theta_{ab},\theta_{bc},\theta_{ac})$ would be equal to $\Pi_i L_i(\theta_{ab})\Pi_j L_j(\theta_{bc})\Pi_k L_k(\theta_{ac})$. If we were to consider n loci in this way the overall likelihood could be written in term of the n-1 map distances, or map locations, given a certain map function without facing the overparameterisation problem.

The first major investigation in multipoint linkage analysis is concerned with questioning linkage between a new locus and a predetermined linkage group of n-1 loci. Given that the n-1 loci occupy specific ordered sites $w_1, w_2, \ldots, w_{n-1}$, which are assumed to be known without error, Morton (1978) introduced the following test to test whether an n^{th} locus is a part of that linkage group. His test was based on two loci data (or pairwise data) between the i^{th} and the n^{th} locus, expressed in the form of lod scores $z(w_n, w_i)$ for i=1,2,...n-1. The total lod score $Z(w_n)$ accumulated over the n-1 markers would be equal to $\Sigma_i z(w_n, w_i)$ provided that all pairwise data comes from independent families. To write the lod score in terms of the map locations he suggested the choice of Rao et al map function with p=0.35. Locus n is asserted to be part of the linkage group if $Z(\hat{w}_n)>3$, where $\hat{w}_n$ is the MLE of w_n.

The simplicity of this test depends critically on the assumption that all parameters except one ,w_n ,are specified with negligible error. This situation may arise in practice when the n-1 loci are test markers loci with map location estimated with accuracy from a large number of panel control families and the n^{th} locus is a disease locus, or any other rare genetic trait. Since the data for the disease locus will generally be much more limited than the test markers, we may assume that the genetic location of the latter are known exactly (Lathrop et al 1984).

If a family pedigree data is informative for more than two loci, then the derivation of the likelihood should be carried out as explained in section(1.6.1), (i.e) taking into account all information about multipoint recombination events. In practice, however, sometimes the likelihood, or lod scores, for three loci data or more is calculated as if derived from independent two

loci data, (i.e) taking $L(\theta_{ab},\theta_{bc},\theta_{ac})=L(\theta_{ab})L(\theta_{bc})L(\theta_{ac})$. Lathrop et al (1984,1985) discussed the relative efficiency between two point and true three point analysis in terms of the precision of the estimated recombination fractions. As they rightly stated, the relative efficiency of three loci and two loci linkage estimates of recombination fractions, which is defined as the inverse of the ratio of the corresponding variances, depends on the true recombination fraction, the mode of inheritance and the type of the family data. This led them to study the efficiency under various combinations of the above variables, as well as comparing the efficiency when making different assumptions about c in the three point analysis. They found that, in general, there is usually some gain in efficiency when using three point analysis against two point analysis and that this gain is larger if no interference is assumed as opposed to estimating c when the true c is equal to 1. Also the gain becomes substantial, when estimating θ_{ab} or θ_{bc}, if the sites of the flanking loci are known without error and c is assumed equal to 1. Actually this last result, had spread some doubt in our mind about our understanding of their definition of the log likelihood under the pairwise analysis. Given a phase known triple backcross with three codominant alleles, for example, the multipoint log likelihood will factorize into two independent log likelihoods if c=1 ((i.e) $l(\theta_{ab},\theta_{bc}) = l_1(\theta_{ab})l_2(\theta_{bc})$) which is, in our understanding equivalent to the log likelihood given a pairwise analysis. Nevertheless their results directed them to suggest the use of multipoint data and analysis under the Haldane map function in deriving a test for detecting linkage between an n^{th} locus and the predetermined n-1 loci, instead of using pairwise data. Under the Haldane map function no problem of

overparameterisation will be faced, if (n-1)>2, and the likelihood function will be only function of one unknown parameter w_n. Proceeding with the testing problem in a fixed sample size context, they suggested the use of the generalised likelihood ratio test to detect linkage, (i.e) comparing $2\ln(10)Z(\hat{w}_n)$ to a x^2 variate with one degree of freedom.

Another study concerning the efficiency of the two point analysis of three point data has been carried out by Maclean et al(1985), they rightly pointed out that there is no statistical justification of calculating the likelihood of three point data as if coming from independent families, the resulting function will not be a log likelihood, though it could be used for deriving estimates of recombination fractions and map distances. Using simulation, they calculated the relative efficiency of the two point analysis to the three point one by calculating the ratio of the variance of $(\hat{w}_1+\hat{w}_2)$ for sample sizes varrying from 50 to 1000 and over relative true distance, $w_1:w_2$, varrying from 1:1 to 50:1; the true map function used in their study has not been mentioned. Their result showed that, as an average over the mentioned sample sizes and distance ratios, the relative efficiency of the two point analysis to the three point one was greater than 0.95, for the range of distance of practical interest ((i.e) less than 20cMorgan). A recent critical paper, which compare both the pairwise and multipoint analysis when the Haldane map function is assumed, is given by Morton(1988).

The second major investigation in multipoint linkage analysis is the determination of the order of a linked group of loci. With n linked loci, the total number of different gene orders is equal to n!/2, where any sequence of loci is considered the same as that resulting from a reversal of it. Given a certain order the

likelihood of $w_1, w_2, .., w_n$ could be derived either from independent two points and a certain map fuction (Morton 1978) or from multipoint data with the Haldane map function (Lathrop et al 1984). An exact determination of the right order is clearly very difficult to achieve, in practice however, an acceptable strategy is to report the most likely gene orders, (i.e) those orders with the highest maximised likelihood.

The problem of testing different gene orders against each other when three loci data are considered, constitutes a major part of the present study. The reader is referred to chapter 2 and chapter 3 for a detailed assessment of the problem. On the meantime it is worth noticing here that as the dimensionality of the likelihood is the same under the different orders, the usual large sample theory of generalized likelihood ratio testing will not be applicable here.

At the end of this background chapter, I would like to refer the interested reader to Ott's (1985) book, which has been referenced throughout this chapter, for quite a clear and extensive discussion of human linkage problems, as well as referring to Smith's (1986) paper for an interesting summary about the development of human linkage analysis.

CHAPTER TWO: Finding the order of three loci - an introduction

(2.1) Introduction

The aim of the following two chapters is to study the problem of the unknown orders of different loci known to be on the same chromosome. A three loci situation is adopted. Criteria used by the geneticists to identify the right order when the three orders are tested simultaneously are mentioned and then studied in various details.

Given a true order and a certain mating type, data are simulated and then used to model the probabilities of right and wrong order decision, made in the light of the chosen criteria, as functions of the recombination fractions. Results and assessment of these models, when a certain map function is assumed, are given in chapter 3. Also, at the end of the present chapter, we discuss testing two orders only against each other as presented by some other authors who are geneticists.

(2.2)Notation and criteria

For simplicity, a three loci situation with two codominant alleles at each locus and a phase known triple backcross mating have been assumed. Let (A,a), (B,b) and (C,c) be the three pairs of codominant alleles. The order of these loci is unknown, therefore, the following three orders, ABC, BAC and ACB, denoted by O_1,O_2 and O_3 respectively, are possible. In chapter one, a definition of α, β, γ and δ was given for a certain order which was O_1; a general definition could be given as follows, let

α be the probability of a recombination in the first and the second segments.

β be the probability of a recombination in the first but not the second segment.

γ be the probability of a recombination in the second but not the first segment.

δ be the probability of no recombination in any segment.

Also let θ_1 and θ_2 be the recombination fraction in the first and the second segment respectively and θ_{1+2} be the recombination in the first or the second segment but not in both, (i.e) the recombination fraction between the first and third loci. Under this general notation, formula(1.6) and (1.12) will become

$$\left.\begin{aligned} \theta_1 &= \alpha+\beta \\ \theta_2 &= \alpha+\gamma \\ \theta_{1+2} &= \beta+\gamma \end{aligned}\right\} \quad \leftrightarrow \quad \left.\begin{aligned} 2\alpha &= \theta_1+\theta_2-\theta_{1+2} \\ 2\beta &= \theta_1-\theta_2+\theta_{1+2} \\ 2\gamma &= \theta_{1+2}+\theta_2-\theta_1 \end{aligned}\right\} \quad (2.1)$$

$$\leftrightarrow \quad \left.\begin{aligned} \alpha &= c\theta_1\theta_2 \\ \beta &= \theta_1(1-c\theta_2) \\ \gamma &= \theta_2(1-c\theta_1) \end{aligned}\right\} \quad (2.2)$$

Also restriction(1.11) and (1.13) will become

$$\left.\begin{aligned} 0 < \theta_1 &< 0.5 \\ 0 < \theta_2 &< 0.5 \\ \theta_1+\theta_2-2\theta_1\theta_2 < \theta_{1+2} &< \text{Min}[0.5;\theta_1+\theta_2] \end{aligned}\right\} \quad (2.3)$$

$$\left.\begin{aligned} 0 < \theta_1 &< 0.5 \\ 0 < \theta_2 &< 0.5 \\ \text{Max}\left[0;\ \frac{\theta_1+\theta_2-0.5}{2\theta_1\theta_2}\right] < c &\leqslant 1 \end{aligned}\right\} \quad (2.4)$$

Now, let

r_1 be the number of offsprings recombinant in segment AB and segment BC.

r_2 be the number of offsprings recombinant in segment AB but not in segment BC.

r_3 be the number of offsprings recombinant in segment BC but not in segment AB.

r_4 be the number of offsprings non recombinant in any segment.

Under the first order, O_1, the first and second segments correspond to segment AB and segment BC respectively. Therefore the probability corresponding to r_1, r_2, r_3 and r_4 will be α, β, γ and δ respectively and it follows from (1.14) that

$$r_1\ r_2\ r_3\ r_4 \sim \text{Mult}(n;\ \alpha\ \beta\ \gamma\ \delta) \qquad (2.5,a),$$

Under the second order, O_2, the first and second segment will correspond to segment BA and segment AC respectively. A recombination in AB and BC means that there is an odd number of crossover between the 1^{st} and 2^{nd} loci (A and B) and an odd number of crossovers between the 1^{st} and 3^{rd} loci (B and C) and thus an even number of crossovers between the 2^{nd} and 3^{rd} loci (A and C). It follows then, that the probability corresponding to r_1 will be the probability of a recombination in the first but not the second segment. Proceeding in this way, it is easy to see that

$$r_1\ r_2\ r_3\ r_4 \sim \text{Mult}(n;\ \beta\ \alpha\ \gamma\ \delta) \qquad (2.5,b).$$

By analogy, under the third order, O_3

$$r_1\ r_2\ r_3\ r_4 \sim \text{Mult}(n;\ \gamma\ \beta\ \alpha\ \delta) \qquad (2.5,c),$$

As a summary, a two by two table showing the number of offsprings corresponding to the pattern of recombination in the first segment by the pattern of recombination in the second segment, for each order, is produced. Table(2.1,a), (2.1,b) and (2.1,c) correspond to the above table when the assumed order is O_1, O_2 and O_3 respectively. Note that the probability corresponding to the first, second, third and fourth cell of any of these tables are α, β, γ and δ respectively.

Now, let P_i be the probability that order O_i is considered correct given the true order, where i=1,2,3. Also let P_4 be the probability that no conclusion is made with respect to the orders given the true order. If O_1 is the true order then the following

Table(2.1) The number of offsprings recombinant and non-recombinant in two segment formed by the three loci A, B and C.

(a) Under order O_1 (ABC).

		recombination in BC	
		yes	no
In AB	yes	r_1	r_2
	no	r_3	r_4

(b) Under order O_2 (BAC)

		recombination in AC	
		yes	no
In AB	yes	r_2	r_1
	no	r_3	r_4

(c) Under order O_3 (ACB)

		recombination in BC	
		yes	no
In AC	yes	r_3	r_2
	no	r_1	r_4

three probabilities, the probability of right decision, the probability of wrong decision and the probability of no conclusion, denoted by P_1, $P_W=P_2+P_3$ and P_4 respectively, will be of interest.

Ott(1985 page183) mentioned two criteria for determining the most likely gene order, both depend on the maximized likelihood function under the three rival hypotheses which is denoted by μ_i for order O_i, (i.e) $\mu_i= L(\hat{\theta}_i|O_i)$, where $\hat{\theta}_i$ is the MLE of $\underline{\theta}=(\theta_1\ \theta_2\ \theta_{i+2})^T$ under order O_i. He stated that a difference between the maximized achieved $\log_e$ (ln) likelihood under the two rival hypotheses is sometimes taken to be relevant when it exceeds the value of two units in lnL.

Let $\lambda_{ij}= \ln \frac{\mu_i}{\mu_j}$ (2.6),

Under the 1st criterion O_i is significantly more likely than O_j if $\lambda_{ij} >2$. Therefore:

$P_i= \text{Prob}\{ \lambda_{ij} >2 \text{ and } \lambda_{is} >2 \mid \text{true order}\}$ (2.7),

where i=1,2,3 and {i,j,s}={1,2,3}.

Another approach is to write down the approximate posterior probability, Y_i, of the i^{th} order as

$$Y_i = \frac{\pi_i\mu_i}{\sum_i \pi_i\mu_i} \quad (2.8),$$

The second criterion has not been mentioned explicitly in Ott(1985), but probably, O_i will be considered the most likely order among all rival hypotheses if Y_i is greater than a certain constant, when $\pi_1=\pi_2=\pi_3=(1/3)$, and therefore

$P_i= \text{Prob}\{ Y_i >\text{const} \mid \text{true order}\}$ (2.9),

where i=1,2,3.

(2.3) Distributional study of the λ_{ij}(s)

As seen in section(2.2) P_1, P_W and P_4 will depend on either the

bivariate distribution of $(\lambda_{ij}\ \lambda_{is})$ or on the distribution of Y_i. This section will be concerned with the investigation of some aspects of the bivariate distribution of $\underline{\Lambda} = (\lambda_{ij}\ \lambda_{is})^T$, such as the expected value, $E(\underline{\Lambda})$ and the covariance matrix, $Var(\underline{\Lambda})$. But the $\lambda_{ij}(s)$ are functions of the maximimized likelihoods under the various gene orders, therefore they are functions of the data vector $\underline{R}=(r_1\ r_2\ r_3)^T$. Now, let

$$\begin{aligned} r_{12} &= r_1+r_2 & 2r_1 &= r_{12}+r_{13}-r_{23} \\ r_{13} &= r_1+r_3 \quad \leftrightarrow & 2r_2 &= r_{12}+r_{23}-r_{13} \\ r_{23} &= r_2+r_3 & 2r_3 &= r_{13}+r_{23}-r_{12} \end{aligned}$$

As r_{12}, r_{13} and r_{23} give a nonsingular transformation of r_1, r_2 and r_3 then for convenience let $\underline{R}=(r_{12}\ r_{13}\ r_{23})^T$ instead.

Taylor expansion can be used to find an approximate expected value and variance of a function of $\underline{R}$, $h(\underline{R})$ as follows

(i) $h(\underline{R})$ can be expanded about $E(\underline{R})$ as

$$\left.\begin{aligned} h(\underline{R}) = h(E(\underline{R})) &+ (\underline{R} - E(\underline{R}))^T \frac{\partial h(E(R))}{\partial \underline{R}} \\ &+ \frac{1}{2} (\underline{R} - E(\underline{R}))^T H(\underline{R}) (\underline{R} - E(\underline{R})) + \ldots \end{aligned}\right\} \quad (2.10)$$

where $H(\underline{R})$ is a matrix whose $(ij)^{th}$ element is $\left.\frac{\partial^2 h(R)}{\partial r_i \partial r_j}\right|_{\underline{R}=E(\underline{R})}$

(ii) By using the first three terms in (2.10), $E(h(\underline{R}))$ will be approximated by

$$\left.\begin{aligned} E(h(\underline{R})) \simeq h(E(\underline{R})) &+ 0 + \\ &+ \frac{1}{2} E\left[(\underline{R} - E(\underline{R}))^T H(\underline{R}) (\underline{R} - E(\underline{R}))\right] \end{aligned}\right\} \quad (2.11)$$

(iii) By using the 1^{st} and 2^{nd} terms in (2.10), $Var(h(\underline{R}))$ is approximated by

$$Var(h(\underline{R})) \simeq 0 + Var(\underline{b}^T\underline{R}) = \underline{b}^T Var(\underline{R})\ \underline{b} \quad (2.12)$$

where $\underline{b} = \frac{\partial\, h(E(R))}{\partial \underline{R}}$

(iv) Similarly we can find an approximate covariance of the two functions $h(\underline{R})$ and $k(\underline{R})$, by using the first two terms in (2.10)

$$\mathrm{Cov}(h(\underline{R})\ k(\underline{R})) \simeq \underline{b}^{T}\ \mathrm{Var}(\underline{R})\ \underline{d} \qquad (2.13)$$

where $\underline{d} = \frac{\partial\, k(E(R))}{\partial \underline{R}}$

Under the general model described in (2.5,a,b,c) the MLE of $\underline{\theta}_i$ will either lie within the feasable region of order O_i or on one of the boundaries. Therefore $\mu_i(\underline{R})$ will have various forms depending on $\hat{\underline{\theta}}_i$, (eg) if the MLE of $\underline{\theta}_1$ lies within the feasable region as described in (2.3) then $\mu_1(\underline{R})$ will have the following form:

$$\mu_1(\underline{R}) = \text{constant} \times \left[\frac{r_1}{n}\right]^{r_1}\left[\frac{r_2}{n}\right]^{r_2}\left[\frac{r_3}{n}\right]^{r_3}\left[\frac{r_4}{n}\right]^{r_4}$$

$$\text{If} \quad \frac{r_{12}}{n} < \frac{1}{2} \quad ; \quad \frac{r_{13}}{n} < \frac{1}{2}$$

$$\text{and} \quad \frac{2r_1+r_{23}}{n} - \frac{2r_{12}r_{13}}{n^2} < \frac{r_{23}}{n} < \mathrm{Min}\left[\frac{1}{2};\ \frac{2r_1+r_{23}}{n}\right]$$

If one of these restrictions is not satisfied then $\mu_1(\underline{R})$ will have a different form. On the other hand, if a certain map function is assumed, the number of independent parameters will be reduced to two instead of three. Also, the third restriction in either (2.3) or (2.4) will not be needed because it will always be satisfied given any map function (see appendix(2.1)). If this map function is any one other than the Haldane then $\hat{\underline{\theta}}_i$ will have to be found numerically. Thus finding μ_i and therefore $E(\underline{\Lambda})$ and $\mathrm{Var}(\underline{\Lambda})$ under either the general model or when a map function other than the Haldane is used will be troublesome.

Given the Haldane map function, it is easy to see from (2.2) that under any order

$$\left.\begin{aligned} \alpha &= \theta_1\theta_2 \\ \beta &= \theta_1(1-\theta_2) \\ \gamma &= \theta_2(1-\theta_1) \end{aligned}\right\} \quad (2.14)$$

therefore under O_1 and from (2.5)(a) and (2.14) it is easy to see that r_{12} is independent of r_{13} where

$r_{12} \sim bi(n\ \theta_1)$

$r_{13} \sim bi(n\ \theta_2)$

Therefore, $E(\underline{R}) = (n\theta_1\ n\theta_2\ n\theta_{1+2})^T$, where $\theta_{1+2} = \theta_1+\theta_2-2\theta_1\theta_2$ and by using the properties of the multinomial distribution of $r_1\ r_2\ r_3\ r_4$, it is easy to find that

$$Var(\underline{R}) = \begin{bmatrix} n\theta_1(1-\theta_1) & 0 & n\theta_1(1-\theta_1)(1-2\theta_2) \\ & n\theta_2(1-\theta_2) & n\theta_2(1-\theta_2)(1-2\theta_1) \\ & & n\theta_{1+2}(1-\theta_{1+2}) \end{bmatrix}$$

where for example,

$$Cov(r_{12}\ r_{23}) = Var(r_2)+Cov(r_1\ r_2)+Cov(r_1\ r_3)+Cov(r_2\ r_3)$$
$$= n\theta_1(1-\theta_1)(1-2\theta_2)$$

Also, $$\mu_1(\underline{R}) \propto \left[\frac{r_{12}}{n}\right]^{r_{12}}\left[\frac{r_{34}}{n}\right]^{r_{34}}\left[\frac{r_{13}}{n}\right]^{r_{13}}\left[\frac{r_{24}}{n}\right]^{r_{24}}$$

If $\frac{r_{12}}{n} < \frac{1}{2}$; $\frac{r_{13}}{n} < \frac{1}{2}$

Similarly, under O_2 and O_3, $\mu_2(\underline{R})$ and $\mu_3(\underline{R})$ could be found. Now given that r_{12}, r_{13} and r_{23} are small enough, (i.e) that each of them is less than (n/2), then:

$$\lambda_{12}(\underline{R}) = \ln \frac{r_{13}^{r_{13}}\ r_{24}^{r_{24}}}{r_{23}^{r_{23}}\ r_{14}^{r_{14}}} \quad \text{and} \quad \lambda_{13}(\underline{R}) = \ln \frac{r_{12}^{r_{12}}\ r_{34}^{r_{34}}}{r_{23}^{r_{23}}\ r_{14}^{r_{14}}}$$

Let $\underline{b}_1 = \frac{\partial\, \lambda_{12}(E(R))}{\partial \underline{R}}$ and $\underline{b}_2 = \frac{\partial\, \lambda_{13}(E(R))}{\partial \underline{R}}$

If the true order is O_1, then

$$\underline{b}_1^T = \left[\begin{array}{ccc} 0 & \ln\frac{\theta_2}{(1-\theta_2)} & \ln\frac{(1-\theta_{1+2})}{\theta_{1+2}} \end{array}\right]$$

$$\underline{b}_2^T = \left[\begin{array}{ccc} \ln\frac{\theta_1}{(1-\theta_1)} & 0 & \ln\frac{(1-\theta_{1+2})}{\theta_{1+2}} \end{array}\right]$$

$$H_{12}(\underline{R}) = \frac{\partial^2 \lambda_{12}}{\partial \underline{R}^2}\Big|_{\underline{R}=E(\underline{R})} = \mathrm{Diag}\left[\begin{array}{ccc} 0 & \frac{1}{n\theta_2(1-\theta_2)} & \frac{-1}{n\theta_{1+2}(1-\theta_{1+2})} \end{array}\right]$$

$$H_{13}(\underline{R}) = \frac{\partial^2 \lambda_{13}}{\partial \underline{R}^2}\Big|_{\underline{R}=E(\underline{R})} = \mathrm{Diag}\left[\begin{array}{ccc} \frac{1}{n\theta_1(1-\theta_1)} & 0 & \frac{-1}{n\theta_{1+2}(1-\theta_{1+2})} \end{array}\right]$$

Using formula (2.11)

$$E(\lambda_{12}(\underline{R})) \simeq \lambda_{12}(E(\underline{R})) + \frac{1}{2} E[(\underline{R} - E(\underline{R}))H_{12}(\underline{R})(\underline{R} - E(\underline{R}))]$$

$$\simeq \lambda_{12}(E(\underline{R})) + 0$$

Similarly, $E(\lambda_{13}(\underline{R})) \simeq \lambda_{13}(E(\underline{R}))$ } (2.15)

Also from formula (2.12)

$$\mathrm{Var}(\Lambda) = \left[\begin{array}{cc} \underline{b}_1^T \mathrm{Var}(\underline{R})\, \underline{b}_1 & \underline{b}_1^T \mathrm{Var}(\underline{R})\, \underline{b}_2 \\ & \underline{b}_2^T \mathrm{Var}(\underline{R})\, \underline{b}_2 \end{array}\right] \qquad (2.16)$$

(2.4)Simulated data and preliminary observations

In the previous section we succeeded in finding an approximation to $E(\underline{\Lambda})$ and $\mathrm{Var}(\underline{\Lambda})$ under the Haldane map function. Although those two measurements are of importance in the exploration of the distribution of $\underline{\Lambda}$, they do not give us any hint about the shape of the distribution. In this section, simulated data, (see appendix(2.2)) will be used to overcome this problem as well as bringing light on the magnitude of P_1, P_W and P_4.

Given that order O_1 is the true order and under the Haldane map function any simulation will depend on the following parameters:

(i)θ_{ab} and θ_{bc}, which we decided to vary between 0.05 to 0.25 in step of 0.05 for each θ respectively.

(ii)n, the total number of offsprings, which we decided to vary

between 5 and 30 in step of 5 and between 40 and 100 in step of 10.

For some of those 325 combinations of the parameters the following has been done:

(a)Simulate I values of r_1, r_2, r_3 and r_4 from the multinomial distribution described in (2.5,a).

(b)For each of the I simulations, find the maximized likelihood function under each order, μ_i, and therefore calculate λ_{ij} and Y_i.

Then, using all I simulated data:

(c)Produce a normal probability plot, and a stem and leaf for the λ_{ij}(s).

(d)Produce a two-dimensional plot of λ_{ij} against λ_{is}.

(e)Produce a ternary diagram -see later- for Y_1, Y_2 and Y_3.

(f)Find the estimated expected value and variance of $\underline{\Lambda}$, from the simulation, and compare it with the approximate $E(\underline{\Lambda})$ and $Var(\underline{\Lambda})$.

Steps (c) and (d) have been done in order to see if there is any ground for a normality assumption about the distribution of $\underline{\Lambda}$ for the value of n used. If this assumption happen to be true, then step (c) should show the normality of each λ_{ij} in the form of a straight line probability plot and a bell shape stem and leaf, whereas step (d) should show the bivariate normality of $\underline{\Lambda}$ in the form of a simulated data scattered in the shape of an ellipse.

From the definition of the Y's in (2.8), it is easy to see that, when $\pi_i=(1/3)$, $Y_1+ Y_2+ Y_3=1$, therefore a convenient way to represent the variability in Y_1, Y_2 and Y_3 would be to use what is called the ternary diagram, as shown in figure (2.1)(a). The triangle shown on this figure with vertices 1,2,3 has the following properties:

(1)It is equilateral and has a unit altitude.

(2)For any point Y in the triangle the perpendiculars Y_1,Y_2 and Y_3 to the side opposite 1,2 and 3 satisfy the following

$$Y_i \geq 0 \quad \text{and} \quad Y_1 + Y_2 + Y_3 = 1 \qquad (2.17)$$

(3)For any data point $Y = (Y_1\ Y_2\ Y_3)$ satisfying (2.17), there is a unique point in the triangle 123 with perpendiculars Y_1, Y_2 and Y_3. Therefore all I simulated data points can be represented in this diagram.

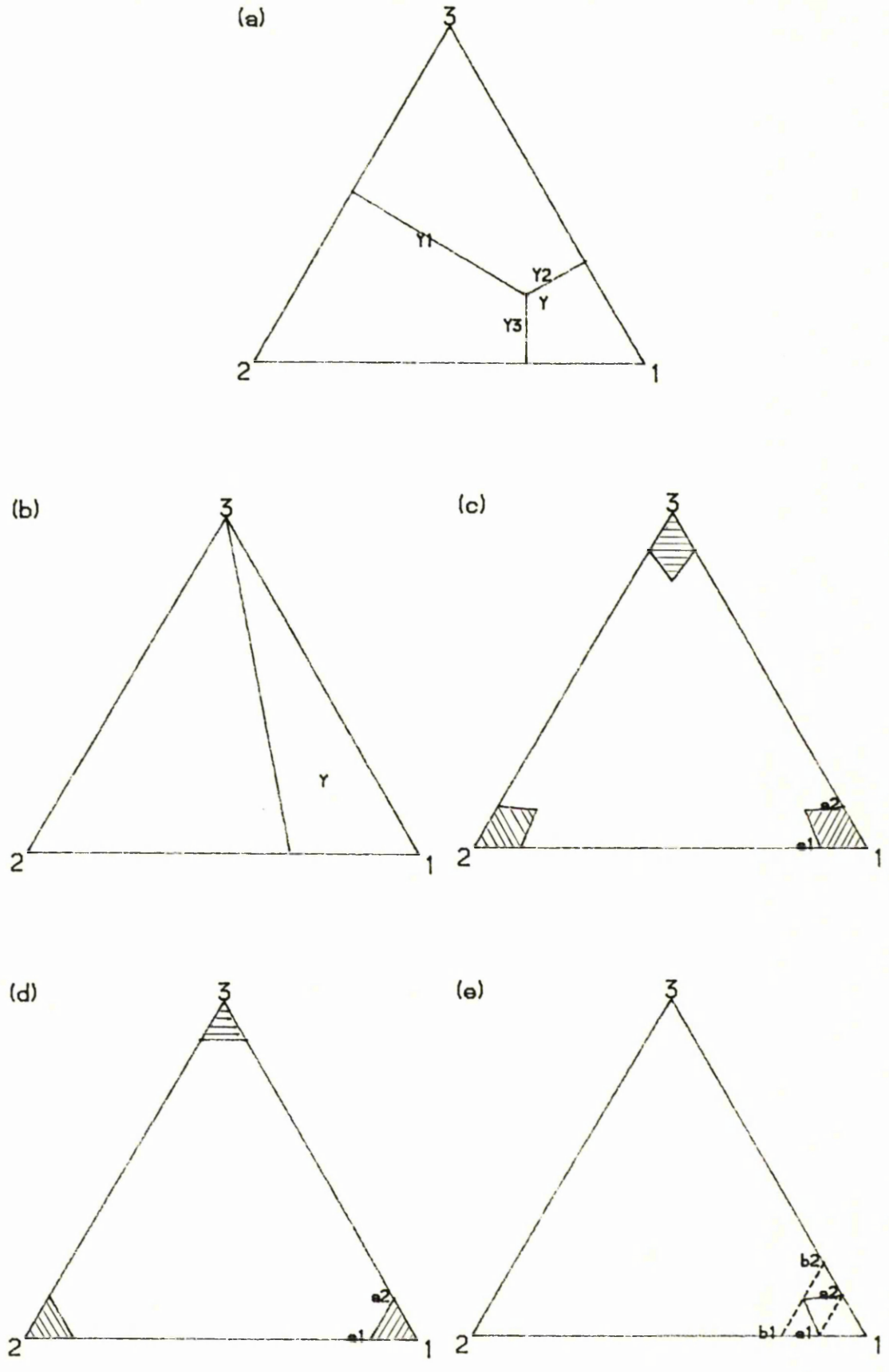
(4)For a vector of points $\underline{Y}$, if two components say Y_2 and Y_3 are in constant ratio, then these points will be represented on a straight line passing through vertex 1. Note that, $\lambda_{12} > 2 \leftrightarrow e^{\lambda_{12}} > e^2 \leftrightarrow (Y_1/Y_2) > e^2$. Therefore, in figure(2.1)(b), the line passing through the vertex 3 and the point a, where

$$a = \left[\frac{e^2}{1+e^2}\ ,\ \frac{1}{1+e^2}\ ,\ 0\right]$$

will divide the triangle into two parts such that a point Y falling in the right side of the triangle will mean that O_1 is significantly more likely than O_2. Dividing the triangle according to all pairwise comparisons between the three orders will produce four areas; three of them near each vertex i and corresponding to considering order O_i as the correct order, the fourth area is the remaining part of the triangle and corresponds to having an inconclusive result -figure(2.1)(c).

(5)For a vector of points $\underline{Y}$, with one component say Y_1, having a constant value, then these points will be represented on a straight line parallel to the line 23. Therefore, dividing the triangle according to the second criterion ((i.e) O_i is considered correct if $Y_i >$ constant) can be represented by figure(2.1)(d) with the same interpretation for the shaded area as in (4).

FIGURE (2. 1)



For convenience, let A(m,l) be the part of the triangle 123 where the order m is considered correct according to criterion l (where m=1,2,3 and l=1,2). Then, for l=2, if

constant$\geq \frac{e^2}{1+e^2}$, A(m,2) will be completely included in A(m,1); whereas if for l=2

constant$\leq \frac{e^2}{2+e^2}$, then A(m,1) will be completely included in A(m,2) -figure(2.1)(e), where a point on a1a2 means that $Y_1 = \frac{e^2}{1+e^2}$ and a point on b1b2 means that $Y_1 = \frac{e^2}{2+e^2}$.

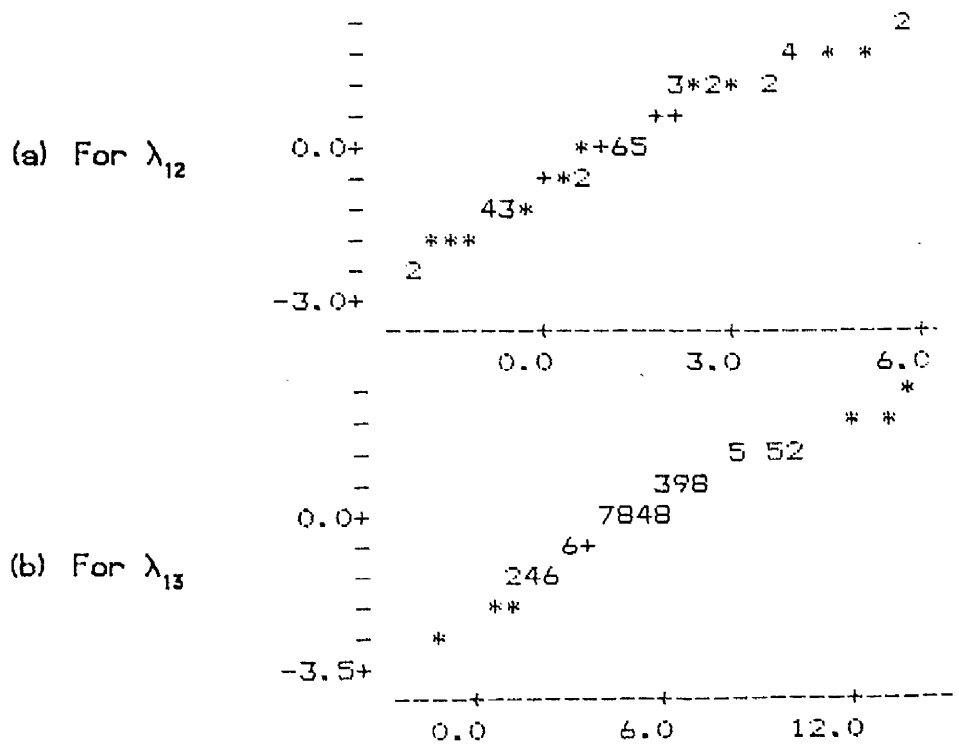
Results:

(i)From our numerous plots, out of which only few are shown here, a bivariate normality assumption for Λ seems to be inadequate, although each λ_{ij} on its own could be seen as normally distributed. Figure(2.2)(a,b,c,d) and (2.3)(a,b,c,d) show the normal probability plot and stem and leaf of λ_{12} and λ_{13} for two different combination of n and $\underline{\theta}$, when I=100. Figure(2.4)(a,b,c,d) shows the two dimensional plot of λ_{13} against λ_{12} for four different combination of n and $\underline{\theta}$, when I=1000, all of them and others, seem to suggest a bivariate distribution of Λ with contours in the form of an arrow head but with slightly different width according to the combination used. Note that, as an example, the six lines $\lambda_{12}= \pm 2$ & $\lambda_{13}= \pm 2$ & $\lambda_{23}= \pm 2$ have been superimposed on figure(2.4)(b) showing the three areas of interest A(m,1), where m=1,2,3.

(ii)On the other hand, studying the ternary diagram seems to suggest, as expected, a different pattern of variability for the various combinations of n and $\underline{\theta}$. For n$\geq$30 and I=100, all I observations tend to lie more and more in the righthand corner of the triangle, which sensibly suggests that as $n \to \infty$, $P_1 \to 1$ irrespective of the value of $\underline{\theta}$ - look at figure(2.5)(a,b,c,d) for

FIGURE (2.2) Probability plot of the λ(s)

For (a) & (b) n=20 θ_{ab}=0.1 θ_{ba}=0.25



For (c) & (d) n=70 θ_{ab}=0.05 θ_{ba}=0.1

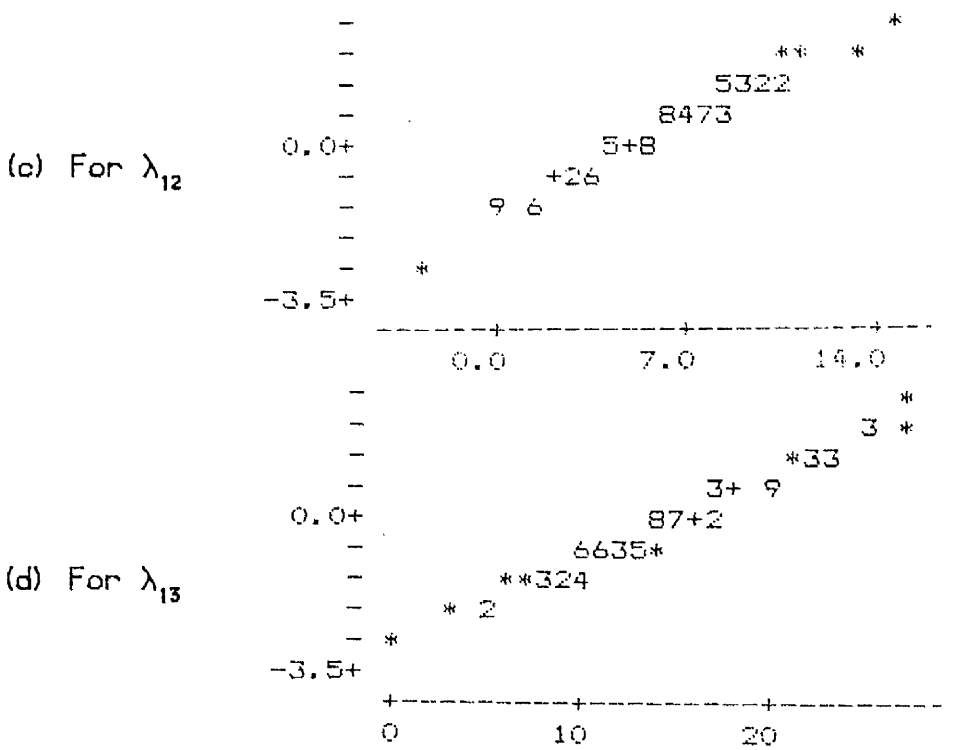


FIGURE (2.3) Stem and Leaf of the $\lambda(s)$

For (a) & (b) n=20 θ_{ab}=0.1 θ_{ba}=0.25

(a) For λ_{12}

Leaf Unit = 0.10

```
 4    -1 9975
 5    -1 2
12    -0 9999775
34    -0 3000000000000000000000
36     0 03
50     0 557889999999999
50     1 222222
44     1 55555777777777777999
24     2 2222222222
14     2 5779
10     3
10     3 557777
 4     4 4
 3     4
 3     5 0
 2     5 77
```

(b) For λ_{13}

Leaf Unit = 0.10

```
 1    -1 2
 1    -0
 3     0 79
 9     1 225999
21     2 222255779999
38     3 455555577777777777
49     4 44444447777
(16)   5 0000334477777777
35     6 000444444445
23     7 22222223
15     8 22222
10     9 44788
 5    10 00
 3    11
 3    12 29
 1    13 7
```

For (c) & (d) n=70 θ_{ab}=0.05 θ_{ba}=0.1

(c) For λ_{12}

Leaf Unit = 0.10

```
 1    -2 6
 1    -1
10    -0 000000000
10     0
22     1 666777899999
30     2 12222246
36     3 558888
47     4 11344557777
(21)   5 11222225555589999999
32     6 3377
28     7 114446666
19     8 0122227789
 9     9 355
 6    10 668
 3    11 4
 2    12
 2    13 3
 1    14 5
```

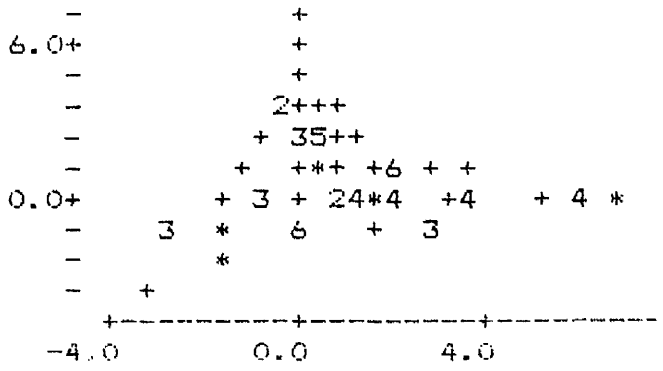
(d) For λ_{13}

Leaf Unit = 1.0

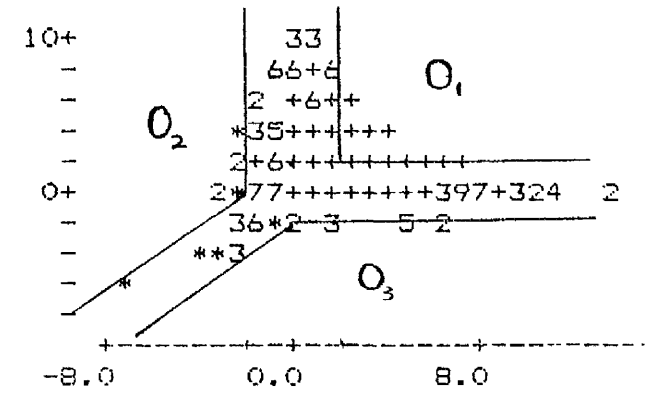
```
 1     0 0
 2     0 2
 5     0 455
 6     0 6
14     0 88888999
30     1 000000000011111
41     1 33333333333
(16)   1 4445555555555555
43     1 66666677777777
29     1 888888899999999
12     2 111
 9     2 2223
 5     2 555
 2     2 66
 0     2
```

FIGURE (2.4) λ_{13} against λ_{12}

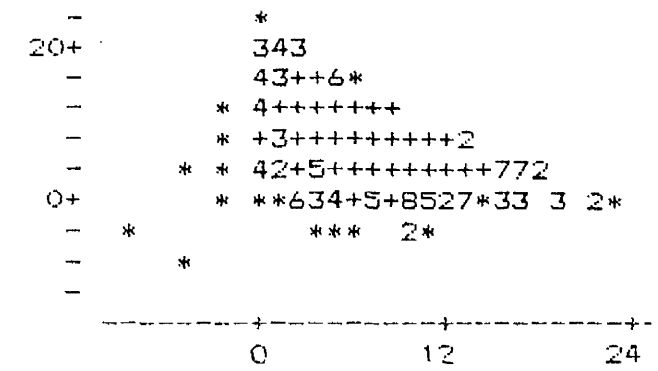
(a) n=10 θ_{ab}=0.1 θ_{bc}=0.2



(b) n=20 θ_{ab}=0.2 θ_{bc}=0.2



(c) n=50 θ_{ab}=0.1 θ_{bc}=0.1



(d) n=100 θ_{ab}=0.05 θ_{bc}=0.25

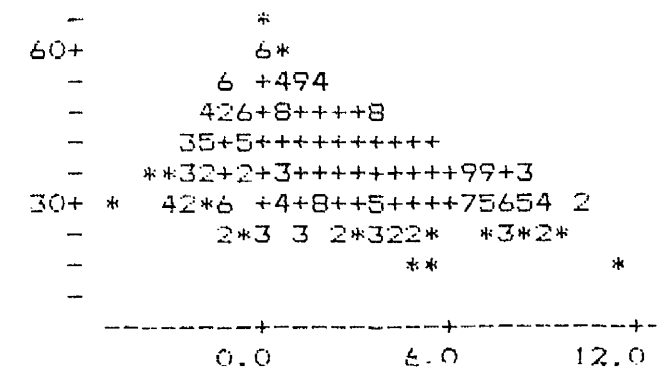
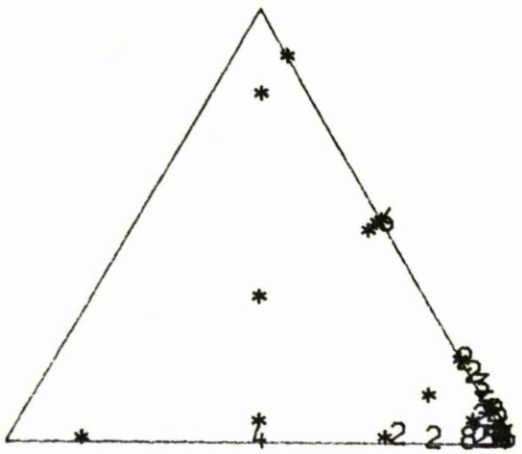
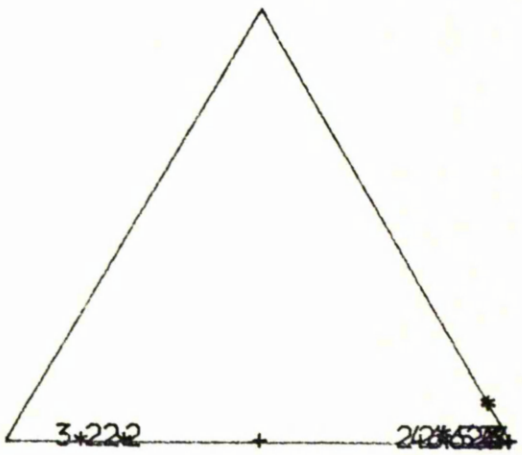


FIGURE (2.5) Ternary, for n=30

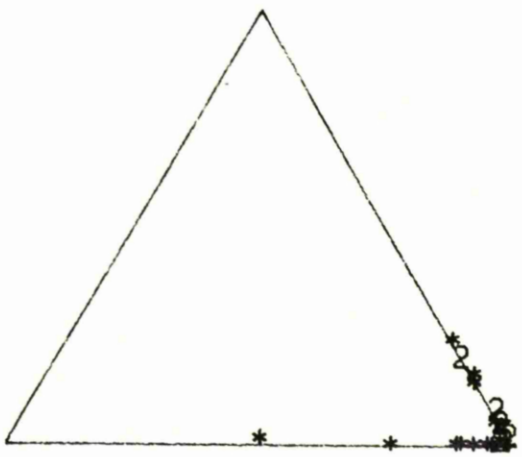
(a) n=30 θ_{ab}=0.1 θ_{bc}=0.1



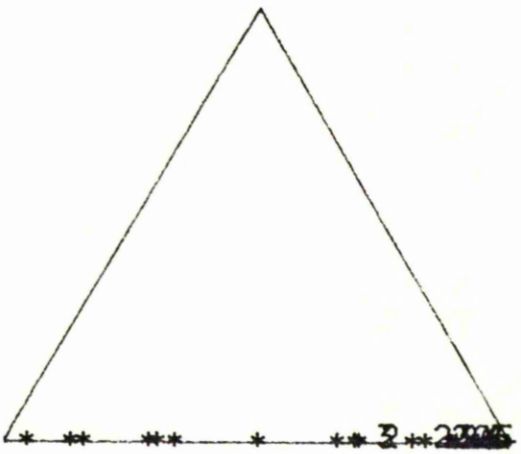
(b) n=50 θ_{ab}=0.05 θ_{bc}=0.15



(c) n=70 θ_{ab}=0.2 θ_{bc}=0.2



(d) n=100 θ_{ab}=0.1 θ_{bc}=0.25



n=30,50,70,100 respectively and when $\underline{\theta}$ varies as shown in the figures. For n<15 a high percentage of the observations seems to lie on either the point c or the segments cc1 or cc2 shown in figure(2.6), suggesting that either $\lambda_{12}=\lambda_{13}=\lambda_{23}$ or $\lambda_{12}=1$ or $\lambda_{13}=1$ respectively, (i.e) meaning that the amount of information held by the likelihood becomes insignificant as n becomes smaller -look at figure(2.7)(a,b) for n=5,10 respectively. For the remaining values of n, (i.e) 15≤n≤25 the 100 observations were more scattered inside the triangle with dependency on $\underline{\theta}$. Therefore, we concluded that in this range of n a more detailed study of P_1, P_W and P_4 as a function of $\underline{\theta}$ should be carried out which is going to be pursued in chapter 3 -look at figure(2.8)(a,b) for n=15,20 respectively.

(iii)Perhaps, it is worth mentioning, as well, that 90%, out of a 100 combinations of n and $\underline{\theta}$, of the approximate $E(\lambda_{ij})$ as calculated from (2.15) have fallen within a 95% standard confidence interval for $E(\lambda_{ij})$ of the form

$$\left[\hat{E}(\lambda_{ij}) \pm 1.96 \left[\frac{\hat{var}(\lambda_{ij})}{100}\right]^{0.5}\right]$$,where $\hat{E}(\lambda_{ij})$ and $\hat{var}(\lambda_{ij})$ are the sample mean and variance calculated using the I=100 simulated observations. The approximate variance $Var(\lambda_{ij})$ as calculated from (2.16) did much worse than that, 20% and 65% of the approximated variances have fallen within a 95% standard confidence interval for $Var(\lambda_{ij})$ of the form

$$\left[\frac{99\ (\hat{Var}(\lambda_{ij}))}{\chi^2(99,0.975)} \quad ; \quad \frac{99\ (\hat{Var}(\lambda_{ij}))}{\chi^2(99,0.025)}\right]$$ when n<40 and n≥40

respectively. This is probably partially due to the approximation of (2.16) and partially due to the sensitivity of the above interval to the normality assumption of the λ_{ij}(s) which would also, perhaps, explain the difference in the performance of the

FIGURE (2. 6)

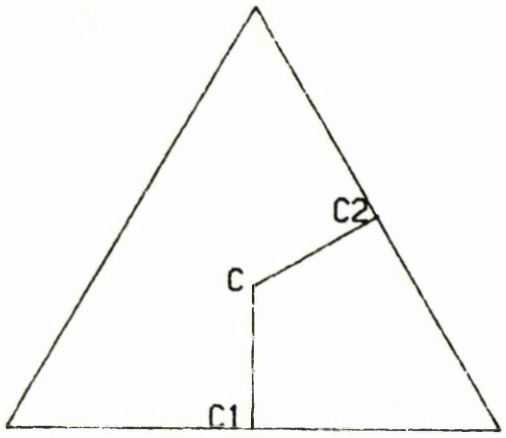


FIGURE (2. 7) Ternary, for n<15

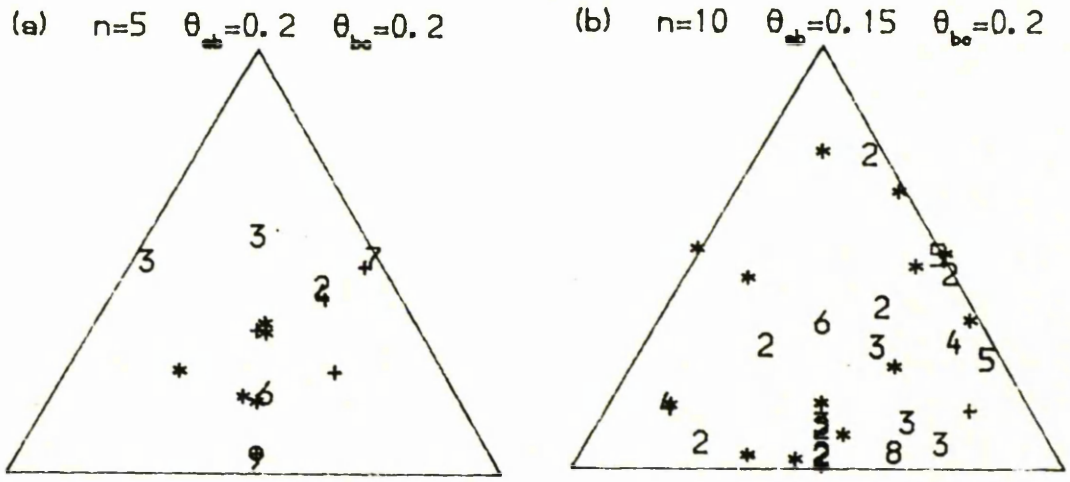
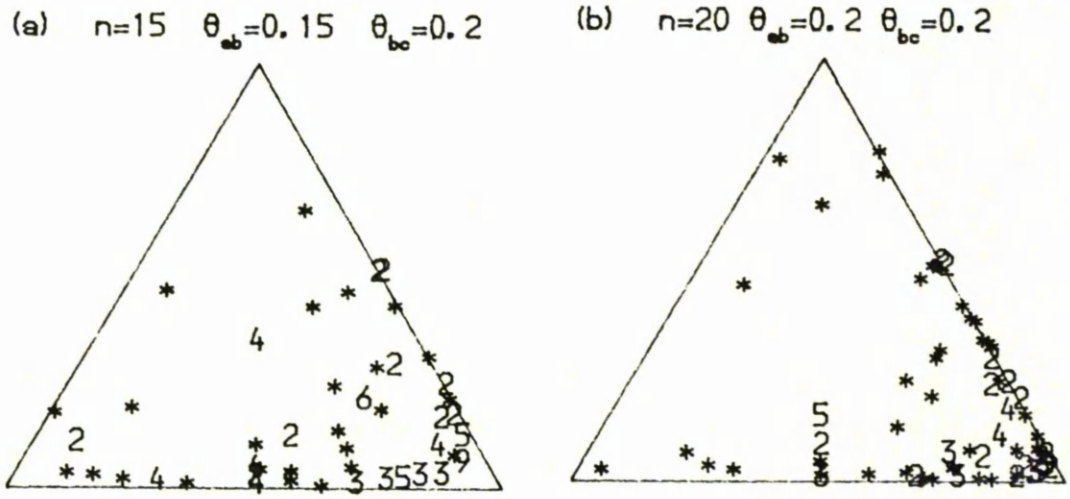


FIGURE (2. 8) Ternary, for 15<n<25



interval for $n<40$ and $n\geq 40$.

(2.5)Symmetry of the P's about true θ_{ab} θ_{bc}

A detailed study of the P's as function of the θ's is going to be carried out in the next chapter, but in order to simplify this forthcoming study, in this section, we are concerned with some aspect of these functions. Our aim is to prove that, for the true order ABC, a certain map function f(.) and for each $P_i(\theta_{ab},\theta_{bc})$, the following is true

$P_i(x,y) = P_i(y,x)$ where $i=1,w,4$.

The general and full proof is given in appendix(2.3), here we are mainly concerned with the main idea and concept of the proof. Actually, the proof depends heavily on the concept that a certain gene order is indistinguishable from its inverse. So that for P_1, which is the probability of deciding that order O_1, ABC, is the right order given it is true, and which therefore depends on the true recombination fraction x and y as the first and second recombination respectively, we will find that by interchanging the values of x and y then this will mean that if a certain observation would have led us previously to decide that order O_1 was the right order then after the interchange it will probably lead us to believe that the inverse of this order, (i.e) CBA, is the right one.

So that

$$P_1(y,x) = P_1(x,y) \qquad (2.18)$$

As for P_2 which is the probabiliy of deciding that O_2, BAC, is the right order given that O_1 is true, we will find that this probability will depend on the true values of x and g(x,y) as the first and second recombination respectively, where $g(x,y)=f^{-1}(f(x)+f(y))$. Similarly for P_3 (concerning order ACB), we will find that this probability will depend on the true values

of g(x,y) and y as being the first and second recombination respectively. Now by just interchanging the values of x and y, we will probably find that an observation that would have previously led us to decide on order O_2 would then lead us to decide on the inverse order of O_3, (i.e) BCA, so that:

$$P_2(y,x) = P_3(x,y)$$

similarly, it is easy to see that

$$P_3(y,x) = P_2(x,y)$$

so that

$$P_W(y,x) = P_2(y,x) + P_3(y,x) = P_W(x,y) \qquad (2.19)$$

Thus from (2.18), (2.19) and from the definition of P_4

$$P_4(y,x) = P_4(x,y) \qquad (2.20)$$

(2.6) Testing two orders only

With three loci situation, Lathrop et al (1987) derived and compared three tests, based on testing the maximum likelihood gene order against one of the orders O_1, O_2 or O_3. The three tests were defined as follows:

$$R_i = \frac{L(\hat{\underline{\theta}})}{L(\underline{\theta}^*)} \qquad \text{for } i=0,1,2$$

Where $\hat{\underline{\theta}}$ is the MLE of $\underline{\theta}$ for the ML order and $\underline{\theta}^*$ is the MLE of the tested order. The difference between R_0, R_1 and R_2 lies in the difference made about the assumption concerning the interference. For R_0, no assumption about interference is made but the constraint of the recombination fraction between the flanking loci being greater or equal to the maximum recombination fraction between the adjacent loci were taken into account, (i.e) in our notation, for R_0 only assume that $\theta_{1+2} > \text{Max}(\theta_1, \theta_2)$. For R_1, the more powerful constraint which assumes positive interference was applied, (i.e) $c<1 \rightarrow \theta_{1+2} > \theta_1 + \theta_2 - 2\theta_1\theta_2$. For R_2, lack of interference was assumed, (i.e) taking c=1 (Haldane map function)

or $\theta_{1+2}=\theta_1+\theta_2-2\theta_1\theta_2$. For any of the three tests the further constraints which assume that any recombination fraction must lie within the range of [0.0,0.5] were also taken into account.

A sensible rejection region would be of the form $R_i>t$. In their paper, they adopted two strategies by which they chose the critical value t_α which would correspond to a certain significance level α. As they are only testing two orders against each other, the significance level is defined as usual as the probability of rejecting the tested order given it is true. As this probability is a function of the true $\underline{\theta}$, their first strategy, which they called the Least favourable strategy, was to choose t_α such that

$$\text{Prob}\{R_i>t_\alpha|\underline{\theta}_L\} = \alpha$$

where $\underline{\theta}_L$ is a certain value of the vector $\underline{\theta}$ which is least favourable in the sence that:

$$\text{Prob}\{R_i>t_\alpha|\underline{\theta}\} \leqslant \text{Prob}\{R_i>t_\alpha|\underline{\theta}_L\}$$

The second strategy, which was called the adaptive strategy, was to choose t_α such that

$$\text{Prob}\{R_i>t_\alpha|\underline{\theta}^*\} = \alpha$$

where $\underline{\theta}^*$ is defined as in the previous paragraph.

Using a triple backcross mating and for a sample of size N=10, 15, 20 and 25 , Lathrop et al produced tables of t_α which correspond to α=0.05, 0.025, 0.01, 0.005 and 0.001 under both strategies. But as they stated, under the adaptive method the significance level α will depend on the order constraint estimates $\underline{\theta}^*$. This led them to produce two critical values for this strategy and suggested that if a data point produced a likelihood ratio R_i larger than the first critical value then the tested hypothesis should be rejected, but if it is smaller than the second critical value then the tested hypothesis should not

be rejected, otherwise if an intermediate result occurs a numerical evaluation of the significance level would be required.

Note that, these tests could be compared to the earlier proposed one in this chapter as follows. The first thing to notice is that the earlier test was concerned with simultaneous testing of the three possible orders whereas Lathrop et al's tests are concerned with testing one order against the ML order. Actually, as the three hypotheses of the three orders O_1, O_2 and O_3 are not nested within each other, then we think that simultaneous testing will be more appropriate, because by doing so we will be able to conclude that either one of the orders is true or that the result was inconclusive, whereas by only testing two orders, as Lathrop et al suggest, no conclusive result as far as the order is concerned will be reached, one order will either be rejected or not. Actually by only testing two orders, the problem remains trapped within the concept of significant and insignificant result which would suit better a null hypothesis nested within a general alternative. Nevertheless, if we consider n loci situation instead of three then using simultaneous testing will certainly lead more and more to an inconclusive result and therefore for this situation rejecting as many orders as possible will be advantageous. Now for a certain data point $\underline{r}$, only one ML order is possible, for the sake of the discussion, let this order be ABC. Under Lathrop et al's tests, the other tested order could either be ABC, BCA or ACB. Given each of these tested orders respectively, the likelihood ratios R_i would correspond to 1, λ_{12} or λ_{13}, in our earlier notation, when the constraint proposed for R_i is taken into account when calculating λ_{12} and λ_{13}. The second point to notice is that, in the earlier test only one critical

value is proposed and as simulteaneous testing is performed the assessment of the test will be based on the performance of the three probabilities P_1, P_W and P_4. On the other hand, Lathrop et al assessed their test by using the usual notion of the size and the power of the test. The size was defined as the sum of the probabilities of those outcomes in the rejection region given that the tested order is the true order, otherwise, if the tested order is different than the true order then this probability defines the power of rejecting the tested order. The definition of the power is a bit unclear but our understanding would be as follows. Let the true order be, ABC, then both the power and the size are defined as follows:

$$\text{Prob}\{R_i > t_\alpha \mid \text{true order ABC}\}$$

This probability is equal the size if the tested order (null hypothesis) against the ML order was ABC. Whereas if the tested hypothesis was BAC, for example, then this probability is equal to the power of rejecting BAC. So that in our notation, <u>firstly</u> if the null hypothesis is ABC:

$$\text{size} = \text{Prob}\{\text{all } \underline{r} \text{ such that the ML order is } O_2 \text{ and } \lambda_{21} > t_\alpha \ \underline{\text{or}} \ \text{all } \underline{r} \text{ such that the ML order is } O_3 \text{ and } \lambda_{31} > t_\alpha \mid \text{ABC}\}$$

<u>Secondly</u>, if the null hypothesis is BAC then:

$$\text{power} = \text{Prob}\{\text{all } \underline{r} \text{ such that the ML order is } O_1 \text{ and } \lambda_{12} > t_\alpha \ \underline{\text{or}} \ \text{all } \underline{r} \text{ such that the ML order is } O_3 \text{ and } \lambda_{32} > t_\alpha \mid \text{ABC}\}$$

Lathrop et al calculated the size and the power of rejecting BAC, for the critical level of 0.05, when the true order was ABC and for some combination of θ_{ab}, θ_{bc} and when the true c was equal 1, the Kosambi level or 0 respectively. When the true c=1, they found that R_0 and R_2 gave the lowest and largest power respectively for most of the cases under both strategies. Also the actual size of any of the tests, R_i, rarely exceeded 0.05,

actually the maximum size occured for R_2 when $\theta_{ab}=\theta_{bc}=0.1$ and was equal 0.055. When interference was present ,(i.e) true c either at the Kosambi level or equal 0, the power of the tests, R_i, became greater. In particular, the power using R_2, which assumes no interference, was quite near the power using R_1, which assumes positive interference, and was even, in some cases, more powerful than R_1 when the true c was at the Kosambi level. The test R_0 performed comparatively better when interference was present but was generally the least powerful. The size of the tests on the other hand was in general conservative when interference was present. Lathrop et al concluded that the testing should be performed by assuming either positive or lack of interference and they were slightly in favour of the latter assumption as it simplifies calculation and was found to be extremely robust.

In the final part of their discussion section, they mentioned that appropriate methods for testing gene orders simultaneously remain to be developed. In the earlier part of this chapter one method has been discussed and remains to be assessed, but admittedly this method lacks a general strategy for the choice of the critical values.

CHAPTER THREE: Finding the order of three loci - a simulation study

(3.1)Modelling P_1 P_W P_4, when N=20

In the previous chapter, a general discussion about the problem of unknown order in linkage analysis was made, from which we found that a detailed study about the variability of P_1, P_W and P_4 as functions of θ_1 and θ_2 would be recommended for $15<N<25$, (where capital N denotes the total number of offsprings). In this chapter we decided to carry out this study using a simulation study the main step of which has been as follows:

(i)For N=20, simulate 1000 values of r_1, r_2, r_3 and r_4 given order O_1, (i.e) using the multinomial distribution described in (2.5)(a), for each combination of θ_{1i} and θ_{2j} where θ_{1i} and θ_{2j} vary as follows, from 0.01 to 0.05 in step of 0.01 and from 0.05 to 0.5 in step of 0.05, therefore the total number of steps=I=J=5+9=14 and the total number of combinations= I×J= 196.

(ii)For each simulated $\underline{r}=(r_1\ r_2\ r_3\ r_4)$ find the maximised likelihood function under all possible orders, (i.e) μ_i i=1,2,3, and therefore λ_{ij}. This step is going to depend on the assumed value of c. If c=1, a simple analytical bounded maximization will be performed but if c≠1, (i.e) Kosambi, eq(3) map function..etc, numerical bounded maximization has to be used. For the Kosambi map function we just made use of the Nag library routine, E04VDF, which applies a quasi-Newton algorithm for finding the maximum of a certain function subject to bounds on the variables, as for the maximisation given Eq(3) some other problems had to be solved first, a detailed discussion of the maximisation given Eq(3) is given in the next chapter. Unimodality of the likelihood will be assumed on the basis of contour plots for some $\underline{r}$. Some of these

plots will be provided later within the relevant section. The numerical maximization has been done using a modified Newton algorithm with the aid of the NAG library routines

(iii)Using all the 1000 simulations for each combination (i,j), calculate n_{1ij}, n_{Wij} and n_{4ij} where

n_{1ij} denotes the number of times order O_1 was considered correct

n_{Wij} denotes the number of times order O_2 or O_3 were considered correct.

n_{4ij} denotes the number of no conclusion decisions.

Therefore, the general model describing (n_1 n_W n_4) would be the following multinomial distribution

$$n_{1ij}\ n_{Wij}\ n_{4ij} \sim \text{Multinomial}(1000;\ P_{1ij}\ P_{Wij}\ P_{4ij})$$

where $P_1+P_W+P_4= 1$ for all (i,j).

therefore

$$L(P_1P_WP_4) \propto \prod_{ij} P_{1ij}^{n_{1ij}} P_{Wij}^{n_{Wij}} P_{4ij}^{n_{4ij}} \qquad (3.1)$$

For any combination (i,j) if we reparametrise our two dimensional parameter vector $\underline{\Pi}= (P_1\ P_4)$, as $P_W= 1-P_1-P_4$, to $\underline{Q}= (Q_1\ Q_2)$ where

$Q_1= \frac{P_1}{1 - P_4}$; (i.e) Q_1 is the probability of the right decision given that a decision has been made, and $Q_2 = P_4$, then the distribution of the data vector $\underline{n}= (n_1\ n_4)$ could be rewritten as follows:

$$f(\underline{n}|\underline{Q})= f(n_4|Q_2)\ f(n_1|n_4\ Q_1)$$

Therefore, n_4 is a sufficient statistic for Q_2 and an ancillary for Q_1. Any inference about Q_2 and Q_1 should be based on the marginal distribution of n_4 and the conditional distribution of n_1 given n_4 respectively. The former is a bi(1000 Q_2) and the latter is a bi(1000-n_4 Q_1). Analogously writing the likelihood (3.1) in terms of $\underline{Q}$ will result in a genuine factorisation as

follows:

$$L(\underline{\Pi}) \propto \prod_{ij} \underbrace{\left[\frac{P_{1ij}}{1-P_{4ij}}\right]^{n_{1ij}}\left[\frac{P_{wij}}{1-P_{4ij}}\right]^{n_{wij}}}_{L(Q_{1ij}|\underline{n}_{ij})} \underbrace{P_{4ij}^{\,n_{4ij}}\,(1-P_{4ij})^{1000-n_{4ij}}}_{L(Q_{2ij}|n_{4ij})}$$

Thus, Q_1 and Q_2, the probability of success of a certain binomial distribution, will be modeled independently using the Generalised linear regression technique which is discussed in the next section. Note that the above arguments will be followed to model Q_1 and Q_2 when:

(a)The Haldane is both the true and the assumed map function.

(b)Eq(3) is the true map function and

(b-1) Haldane is the assumed one.

(b-2) Kosambi is the assumed one.

(b-3) Eq(3) is the assumed one.

Step (b) will be performed in order to measure the sensitivity of the result to the assumed map function. Finally the estimates of Q_1 and Q_2 will be used to provide estimates of P_1, P_W and P_4 where

$$\left.\begin{aligned} P_1 &= Q_1(1-Q_2) \\ P_W &= 1- Q_1(1-Q_2)- Q_2 \\ P_4 &= Q_2 \end{aligned}\right\} \quad (3.2)$$

(3.2)Generalised Linear Model

This section introduces a class of regression models for a scalar response data described by McCullagh and Nelder (1983), with emphasis on the special case of logistic regression model which relates the probability of a binary response variable, Y, taking values 0 and 1, to a covariate vector $\underline{X}= (X_1\ X_2\ \ldots X_p)^T$.

In general, a generalised linear model, GLM, for Y could be

defined through the following three properties:

(i) The probability density function of Y could come from any of the exponential family class of distributions, therefore

$$f_Y(y|\gamma,\phi)= \exp\{ [y\gamma - b(\gamma)]/a(\phi) + c(y,\phi) \} \qquad (3.3)$$

where γ is the unknown canonical parameter and ϕ is usually a known parameter and in some cases unknown. By choosing the appropriate functions a(.), b(.) and c(.,.), many commonly known distributions which come from the exponential family could be written in the above form

(eg) if $mY \sim bi(m\ \mu)$ (3.4), then

$b(\gamma)= \ln(1+e^{\gamma})$

$a(\phi)= 1/m$, where $\phi=1$ and

$c(y,\phi)= \ln \binom{m}{my}$

In general, by using the probability density function of Y $f_Y(y|\gamma,\phi)$, the mean and variance of Y can be derived easily from the following well known relations:

$$E\left[\frac{\partial l}{\partial \gamma}\right] = 0$$

$$E\left[\frac{\partial^2 l}{\partial \gamma^2}\right] + E\left[\frac{\partial l}{\partial \gamma}\right]^2 = 0$$

where $l(\gamma,\phi)= \ln f_Y(y|\gamma,\phi)$,therefore

$E(Y)= b'(\gamma)$

$Var(Y)= b''(\gamma)\ a(\phi)$

where a prime denotes differentiation with respect to γ. Thus, the variance of Y is the product of two functions, the first, $b''(\gamma)$, is a function of the canonical parameter γ and hence the mean, $\mu= b'(\gamma)$, while the second, $a(\phi)$, is independent of γ.

(eg) for the binomial distribution described in (3.4)

$$E(Y)= \mu = \frac{\partial}{\partial \gamma} [\ln(1+e^{\gamma})] = \frac{e^{\gamma}}{1+e^{\gamma}}$$

$$Var(Y)= \frac{\partial \mu}{\partial \gamma} a(\phi) = \frac{\mu(1-\mu)}{m}$$

The above derivation refers to a single observation (or distinct covariate combination). When the data consists of a (K×1) response vector $\underline{Y}$ taking values $\underline{y}$, the parameter γ and therefore μ will be replaced by a parameter vector $\underline{\gamma}$ and $\underline{\mu}$ respectively with component γ_k and μ_k corresponding to each y_k.

By that, we end our discussion of the first property of the GLM which defines the random component of the model through the density function of $\underline{Y}$.

(ii) The second property is concerned with the systematic component of the model. We assume the existence of vectors of covariates $\underline{X}_k= (X_{k1}\ X_{k2}\ \ldots\ X_{kp})^T$ of dimension (p×1), which produce linear predictors η_k given by:

$$\eta_k = \underline{X}_k^T \underline{\beta} = \sum_{l=1}^{p} \beta_l X_{kl} \quad , \text{ for } k=1,..K$$

(note that η_k is linear in the unknown parameters $\beta_l(s)$).

(iii) The third property defines a link function, g(.), between the random and the systematic components of the model such that

$$\eta_k = g(\mu_k)$$

where g(.) is any monotonic differentiable function. By relating each μ_k to the parameter vector $\underline{\beta}$ using the above link function the number of parameters in the model will be reduced if p<K.

A special class of link functions known as the canonical links occur when

$$\gamma_k = \eta_k = \underline{X}_k^T \underline{\beta}$$

An advantage of using them is the simplicity in finding sufficient statistics for the unknown parameters $\underline{\beta}$, where

$$L(\underline{\gamma}\ \phi) = L(\underline{\beta}\ \phi)$$

$$= h(\underline{y},\phi)\ \exp \sum_k [y_k \underline{x}_k^T \underline{\beta} - b(\underline{x}_k^T \underline{\beta})]/a_k(\phi)$$

giving the following sufficient statistics for each β_l

$$\sum_k y_k x_{kl} \quad , \text{ for } l=1,2,\ldots p$$

Again for binomial data $0<\mu_k<1$ and therefore the link function g(.) should satisfy the condition that it maps the interval (0,1) onto the whole real line $(-\infty,\infty)$. The following three link functions are commonly used for that purpose:

1- Logit $\quad \eta_k = \ln \frac{\mu_k}{1-\mu_k}$, which is the canonical link.

2- Probit $\quad \eta_k = \Phi^{-1}(\mu_k)$

3- Complememtary ln-ln $\quad \eta_k = \ln(-\ln(1-\mu_k))$

(3.3) Assessment of GLM

A- Measures of discrepancy

Fitting a model to a set of data may be regarded as a way of replacing the data values $\underline{y}$ by a set of fitted values $\hat{\underline{\mu}}$ derived from a model involving a relatively smaller number of parameters. The simplest model, the null model, has one parameter, representing a common μ for all the y_k's. At the other extreme, the full model has K parameters, one per observation or distinct covariate combination, producing a perfect fit to the data.

Measures of discrepancy between the full model, used as a baseline, and an intermediate model with p parameters, known as the current model, may be defined in various ways. Here we are going to be concerned with two such measures denoted by $D(\underline{y};\tilde{\underline{\mu}})$ and $APA(\underline{y};\tilde{\underline{\mu}})$. The former is the well established measure known as the deviance and defined as the scale parameter ϕ multiplied by twice the difference between the maximum ln likelihood achieved

by the full model and that achieved under the current model. Let $\underline{\mu}^*$ and $\tilde{\underline{\mu}}$ be the estimated means under the full and current model respectively with $\underline{\mu}^*=\underline{y}$. Also, if we denote by $\gamma^*=\gamma(\underline{y})$ and $\tilde{\gamma}=\gamma(\tilde{\underline{\mu}})$ the estimates of the canonical parameters under the two models and by taking $a_k(\phi)=\phi/w_k$, then $D(\underline{y};\tilde{\underline{\mu}})$ could be written as :

$$D(\underline{y};\tilde{\underline{\mu}}) = \sum_k 2w_k[y_k(\gamma_k^* - \tilde{\gamma}_k) - b(\gamma_k^*) + b(\tilde{\gamma}_k)]$$

where the $w_k(s)$ are prior weights known in advance.

The second measure which is called the Average Prediction Ability, is introduced in this study and is defined as the square root of the average squared differences between the estimated linear predictors under the full model, $\eta_k^*=g(y_k)$, and the current model $\tilde{\eta}_k=g(\tilde{\mu}_k)$. Therefore,

$$APA(\underline{y};\tilde{\underline{\mu}}) = \left[\sum_k (\eta_k^* - \tilde{\eta}_k)^2/K \right]^{0.5}$$

In the application presented here, (i.e) for binomially distributed data and with $g(\mu)=\text{logit}(\mu)$

$$D(\underline{y};\tilde{\underline{\mu}}) = 2 \sum_k m_k[y_k \ln(y_k/\tilde{\mu}_k) + (1-y_k)\ln(1-y_k)/(1-\tilde{\mu}_k)]$$

$$APA(\underline{y};\tilde{\underline{\mu}}) = \left[\sum_k [\text{logit}(y_k) - \underline{x}_k^T\underline{\beta}]^2/K \right]^{0.5}$$

For such data let $m=\text{Minimum}(m_1, m_2 \ldots m_K)$, then if $m \to \infty$, the distribution of $D(\underline{y};\tilde{\underline{\mu}})$ given the current model is asymptotically χ^2_{K-p}. Therefore for very large samples, the following test statistics

$T.S = \dfrac{D-E(D)}{[Var(D)]^{0.5}}$, should be approximately normal under the current model.

A more detailed inspection of the second measure for such data is needed here. Let $(\eta_k^* - \tilde{\eta}_k)$ be denoted by $PD(y_k)$. Actually each $PD(y_k)$ is equal to the ln ratio of the estimated odds under the

full and current models respectively. Under the full model $\mu_k^*=y_k$, now let $\tilde{\mu}_k=y_k+e_k$, then

$$PD(y_k) = \ln\left[\frac{y_k}{1-y_k} \Big/ \frac{y_k+e_k}{1-y_k-e_k}\right]$$

$$= \ln\left[\frac{y_k}{y_k+e_k} \Big/ \frac{1-y_k}{1-y_k-e_k}\right]$$

This means that the APA measure relates the error made in the estimation of each y_k to its original value. Thus, if for example $e_1=e_2=0.01$ but with $y_1=0.02$ and $y_2=0.5$, then $PD(y_1)$ will be much larger than $PD(y_2)$ reflecting the seriousness of an error of magnitude 0.01 when related to a small observation relative to a larger one.

Finally it is worth mentioning that, when hypothesis testing between two rival models is not of major importance, a simple comparison between the two models on the basis of their APA will be of use. Actually, the APA of a certain model could be seen as a measure of the average error in the prediction of logit(μ_k) made by that model, but unadjusted to the number of parameters assumed by the model.

B-Residual

The second method of assessing the fit of a model is to examine the residuals. The simplest definition of which, in the context of binomial data is the Pearson residual, defined as the difference between the observed and expected counts scaled by the estimated standard deviation of $m_k y_k$, (i.e)

$$r_k = \frac{m_k(y_k-\tilde{\mu}_k)}{[m_k\tilde{\mu}_k(1-\tilde{\mu}_k)]^{0.5}}$$, which are approximately normally distributed for large m_k and μ_k not very near 0 or 1.

(Note that a plot of these residuals against the fitted values for a certain model under consideration is going to be used in

order to assess the various assumptions given by that model).

(3.4)Application

In this section, we are going to apply and discuss in details the prescribed method of fitting and assessing a GLM to the following set of simulated data:

$$n_{4ij} \sim bi(1000\ Q_{2ij}) \qquad (3.5)$$

for i=1,2...,14 & j=1,2,...,14

where n_{4ij} and Q_{2ij} are as defined in section (3.1). For convenience we are going to drop the subscript 4 and 2 in this section. The binomial distribution in (3.5) defines the random component of the model with

$$E(n_{ij}) = 1000Q_{ij}$$

$$Var(n_{ij}) = 1000Q_{ij}(1-Q_{ij})$$

The full model, M_f, defines the systematic component, (i.e) the linear predictor, as follows

$$M_f: \quad n_{ij} = \alpha_{ij} \quad \text{for } i=1,2,...,14 \quad j=1,2,...,14$$

But, as seen in chapter two, $Q_{ij} = Q_{ji}$, therefore a more appropriate model called here the true model, M_t, will define n_{ij} as

$$M_t: \quad n_{ij} = \alpha_{ij} \quad \text{with} \quad \alpha_{ij} = \alpha_{ji} \quad \text{for all } i \neq j$$

Whereas the simplest model, the null model, M_o, will have only one parameter representing a common Q for all the n_{ij}(s), (i.e)

$$M_o: \quad n_{ij} = \alpha$$

Our aim is to find a simple model between M_o and M_t. Two such models are suggested here and are denoted by M_1 and M_2 respectively:

$$M_1: \quad n_{ij} = \alpha + \beta_1(Z_{1i}+Z_{2j}) + \beta_2(Z_{1i}^2+Z_{2j}^2) + \beta_3(Z_{1i}*Z_{2j}) + \beta_4(Z_{1i}^3+Z_{2j}^3) \qquad (3.6)$$

where $Z_{1i} = \text{logit}(\theta_{1i})$; $Z_{2j} = \text{logit}(\theta_{2j})$

$$M_2: \quad n_{ij} = \alpha_i + \beta_{1i} Z_{2j} + \beta_{2i} Z_{2j}^2 + \beta_{3i} Z_{2j}^3 \qquad (3.7)$$

for $i=1,2,\ldots,14$; $j=1,2,\ldots,14$

Note that M_1 is the special case of M_2 when

$$\left.\begin{aligned} \beta_{3i} &= \beta_4 \\ \beta_{2i} &= \beta_2 \\ \beta_{1i} &= \beta_1 + \beta_3 Z_{1i} \\ \alpha_i &= \alpha + \beta_1 Z_{1i} + \beta_2 Z_{1i}^2 + \beta_4 Z_{1i}^3 \end{aligned}\right\} \text{ for } i=1,2,\ldots,14$$

Model M_1 is a simple polynomial form in Z_{1i} and Z_{2j} proposed firstly by the discrete contour plot of $(n_{ij}/1000)$ against θ_{1i} and θ_{2j}, shown in figure(3.1) and from which we can see the shape of the contours forming disturbed ellipses and probably suggesting a polynomial form of $n(\theta_{1i},\theta_{2j})$.(Note that in this figure, and similar ones that are going to be shown later, the vertical scale changes at 0.05). Secondly by the two dimensional plot of $\text{logit}(n_{ij}/1000)$ against Z_{2j}, shown in figure(3.2) and which suggests either a quadratic or a cubic relation between them. (Note that a similar argument could be made for Z_{1i}). Thirdly, M_1 is constructed such that $n_{ij} = n_{ji}$, thus preserving the symmetric property of the $P_i(s)$, i=1,w,4. Finally, it is perhaps worth mentioning that as GLIM package has been used to fit a logistic regression to our data, trial and error had played an important role in introducing model M_1.

Model M_2, on the other hand, which fits a different cubic in Z_2 for each Z_{1i}, has been mainly suggested by the plotting of $\text{logit}(n_{ij}/1000)$ against Z_{2j} for each Z_{1i} - look at figure(3.3) for some examples. Note that for some Z_{1i} the cubic or the quadratic or even the linear term will be insignificant and

When Q_2 is modeled & Haldane is both the true & assumed map function

FIGURE (3.1) Discret contour plot of the data vs θ_1 & θ_2

```
0.50+   ***AA     A   A   A   A   A   A   *   A   *
    -   *AAAA     A   A   A   A   A   A   A   A   A
    -   A*AAA     A   A   A   A   A   A   A   A   A
    -   AAAAA     A   A   A   A   A   A   A   A   A
    -   AAAAA     B   B   B   B   A   A   A   A   A
0.25+   AAAAB     B   C   C   B   B   A   A   A   A
    -   AAABB     C   C   D   C   B   A   A   A   A
    -   AABBB     D   D   D   C   B   B   A   A   A
    -   AABBC     D   D   C   C   B   A   A   A   A

0.050+  ABBBC     C   B   B   A   A   A   A   A   A
    -   ABBBB     B   B   B   A   A   A   A   A   A
    -   ABBBB     B   B   A   A   A   A   A   A   A
    -   AABBB     B   A   A   A   A   A   A   A   *
    -   AAAAA     A   A   A   A   A   A   A   *   *
0.000+
        +---------+---------+---------+---------+---------+
      0.00      0.10      0.20      0.30      0.40      0.50
```

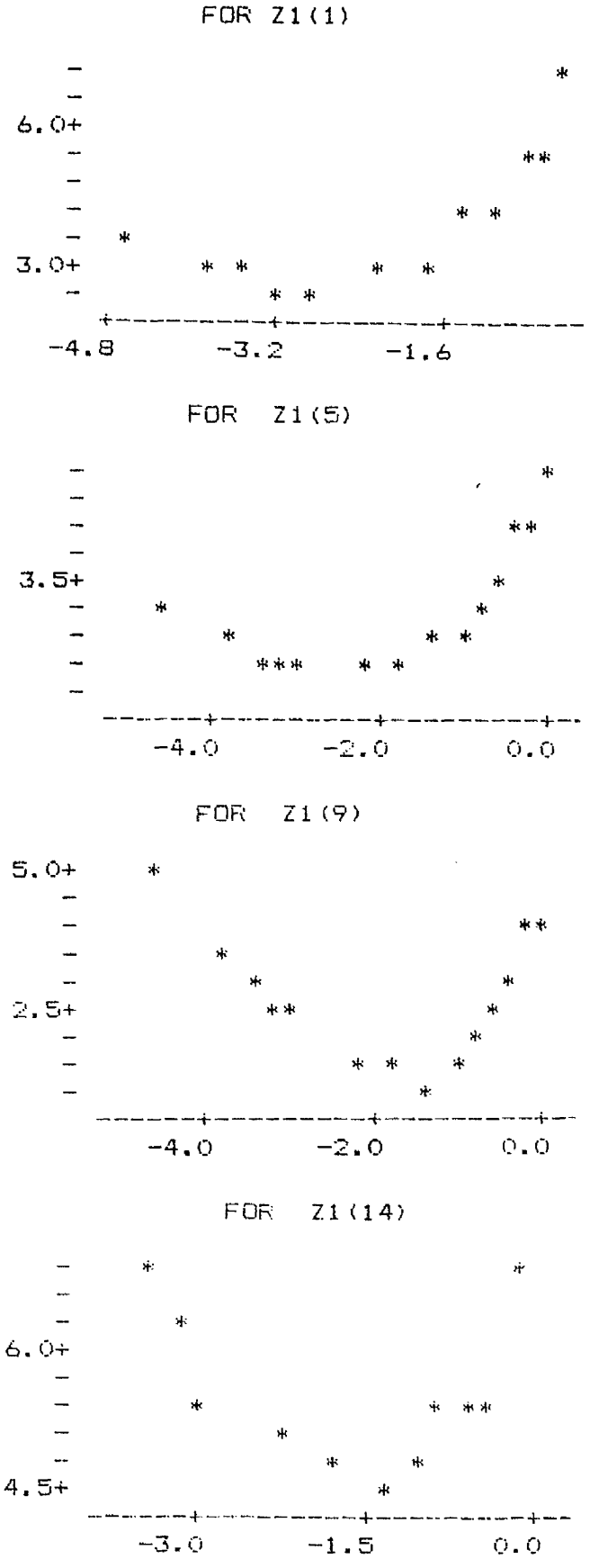
* = Q2=1.000
A = 0.900---0.999 C = 0.700---0.799
B = 0.800---0.899 D = 0.600---0.699

FIGURE (3.2) Logit of the data vs Z_{2j}

```
    -
    -      *    * *                    *22
    -      2    *  2*                2**23
5.0+       *    * 2**    * *     * 263*
    -      2    2 2*2    * * 3 344*63
    -      6    4 22*    3 3 3 3344
    -           4 775    4 4 4 453
    -              *4    5 5 4 3
0.0+
      -------+---------+---------+---
           -4.0       -2.0       0.0
```

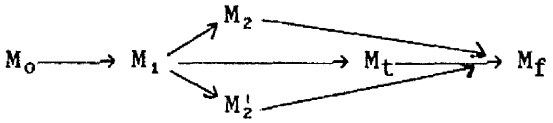
When Q_2 is modeled & Haldane is both the true & assumed map function

FIGURE (3.3) Logit of the data vs Z_{2j} for some Z_{1i}



therefore will not be included in the fit, (i.e) only significant parameters in (3.7) will be included in the analysis. Another model M_2' would fit a different cubic in Z_1 for each Z_{2j}, but as we suspect that this fit will be very similar to that of model M_2, only the analysis given the latter will be executed.

Note that when fitting all the above models,



, where "$\longrightarrow$" means nested within, the logit function has been used as our link function, (i.e) $\text{logit}(\pi_{ij}) = Q_{ij}$.

(3.5)Results

(a)-Haldane is the true and assumed map function

(i)When Q_2 is modeled

Table(3.1) give the relevant summary statistics of the above models, when the Haldane map function is both the true and assumed map function; all entries in the table have been discussed before except for the last column which is just the ratio of the APA of model M_1 to that of the other models. From the table we can see that, despite the substantial drop in the deviance when moving from model M_0 to M_1 to M_2, models M_1 and M_2 still gave significant result when compared to model M_f. Also, we can see that the APA of model M_1, which is equal to 0.31, is 11% and 41% higher than that of model M_2 and M_t respectively.

It is worth mentioning, as well, that 10 observations out of the 196 distinct combinations were equal to "1000". Most of them corresponded to either θ_1 or $\theta_2 \geqslant 0.35$, an area of no practical importance, so to prevent any infinity in the results we decided to just omit them from the analysis.

Table(3.2)(a) give us a more detailed account of the APA of

When Q_2 is modeled and Haldane is both the true and assumed map function.

Table(3.1) A summary of the performance of different models

Model	Deviance	D.f	T.S	APA	$\frac{APA(M_1)}{APA(M_i)}$
M_0	16390.0	185		1.78	
M_1	312.0	181	6.8	0.31	1.00
M_2	164.7	130	2.2	0.28	1.11
M_t	96.8	85	0.9	0.22	1.41

Table(3.2)(a) Individual APA(M_1,i)

i=1,2,...7	0.37	0.27	0.32	0.22	0.20	0.22	0.14
i=8,9,..14	0.11	0.18	0.31	0.43	0.45	0.50	0.35

(The overall APA is equal 0.31)

Table(3.2)(b) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.29	0.19	0.18	0.13	0.11	0.12	0.16
i=8,9,10	0.12	0.09	0.26				

(The overall APA is equal 0.18)

model M_1, this is done by calculating this measure for each θ_{1i}, (i.e) calcutating:

$$APA(M_1;i) = \left[\sum_{j}^{J} (n^{*}_{ij} - \tilde{n}_{ij})^2 / J \right]^{0.5} \quad \text{for each } i=1,2\ldots,14$$

By comparing these individual APA(s) with each others, it seems that the model fits worse for $i \geqslant 11$ and for $i=1$. But since large values of θ_1 and θ_2 are not of great interest it would be better to assess the model's prediction ability when the APA is restricted to the area of θ_1 and $\theta_2 \leqslant 0.3$. Table(3.2)(b) gives the individual APA(s) for the restricted area, showing some drop from their corresponding ones in table(3.2)(a). Actually the overall APA drops from 0.31 to 0.18. In order to see how this 0.18 error in the logit of Q_{ij} is transmitted to Q_{ij} scale we provided the reader with some numerical examples, shown in table(3.3). The first column of this table shows some typical data points $n_{ij}/1000$ which cover the whole range of the data used in the calculation of the 0.18 error. The last two columns show the lower and upper values of $\tilde{Q}_{ij}$ for such data when the 0.18 error is subsequently subtracted and added to $\text{logit}(\tilde{Q}_{ij})$.

Despite the significant deviance of model M_1, it seems to give good result when compared to more general models. Actually with binomial kind of data and where large samples are involved, it is usually expected to have large deviances even for models which fit well, as judged by the closeness of the fitted and actual values - which could be deduced from table(3.3) for model M_1. This happens because with large samples, very small and unimportant deviation from the model can be detected making a significant result more probable. According to the above discussion as well as the extreme simplicity of model M_1, we decided to choose it as our representative of the data, knowing

When Q_2 is modeled and Haldane is both the true and assumed map function.

Table(3.3) Examples showing the errors made by model M_1

$\mu^*=\frac{n_4}{1000}$	Logit(μ^*)	η^*_l= Logit(μ^*) −0.18	η^*_u= Logit(μ^*) +0.18	$\frac{Exp(\eta^*_l)}{1+Exp(\eta^*_l)}$	$\frac{Exp(\eta^*_u)}{1+Exp(\eta^*_u)}$
0.624	0.549	0.369	0.729	0.591	0.675
0.826	1.558	1.378	1.738	0.799	0.850
0.889	2.081	1.901	2.261	0.870	0.906
0.948	2.903	2.723	3.083	0.938	0.956
0.998	6.213	6.033	6.393	0.998	0.998

(The points chosen in the 1st column are the minimum, 1st quartile ,median, 3rd quartile & maximum of the data within the restricted area of θ_1 & $\theta_2 \leq 0.3$).

Table(3.4) MLE and 95% I.E for the parameters of Model M_1

Parameters	MLE	95% Interval estimates
α	6.991	6.840 ; 7.142
β_1	3.804	3.684 ; 3.923
β_2	1.605	1.550 ; 1.660
β_3	-0.419	-0.438 ; -0.400
β_4	0.145	0.138 ; 0.152

that it is not the best choice as judged by the deviance but an approximate one as juged by the APA(s). Maximum likelihood estimates of its unknown parameter vector $\underline{\beta}$ and their corresponding 95% confidence intervals are shown in table(3.4). By using those estimates a contour plot of $\tilde{Q}_2$ ($\tilde{P}_4$) against θ_1 and θ_2 is produced and shown in figure(3.5). Actually this figure is just the estimated continuous version of figure(3.1). Figure(3.4) is just a plot of the Pearson residuals against the fitted values of model M_1, no general trend can be deduced from the plot although we can see that a big cluster of the fitted values points are situated near 1, thus making the interpretation of the plot more difficult as it is expected that at this end the distribution of the residuals will be markedly skewed.

(ii)When Q_1 is modeled

As seen in section(3.1) the conditional distribution of n_1 given n_4 is bi(1000-n_4 Q_1). In this section, for convenience, let for any combination (i,j) , m_{ij} be the number of conclusive decisions out of the 1000, n_{ij} be the number of right decision and Q_{ij} be the probability of a right decision given that a decision has been made, then

$n_{ij} \sim bi(m_{ij}\ Q_{ij})$ for i=1,2,..,14 and j=1,2,..,14.

But, as the same line of analysis which was described in section(3.3) will apply throughout this result section, let us first summarize the main common steps of the analysis:

(A)Four different models are usually tested against M_f

1- The null model M_0.

2- Model M_1, suggested by two figures, called in general figure(3.1,m,s,l) and figure(3.2,m,s,l) respectively, as well as trial and error in GLIM. Figure(3.1,m,s,l) is the discrete contour plot of the data against θ_{1i} & θ_{2j}, whereas

When Q_2 is modeled & Haldane is both the true & assumed map function

FIGURE (3.4) Pearson residuals vs the fitted values

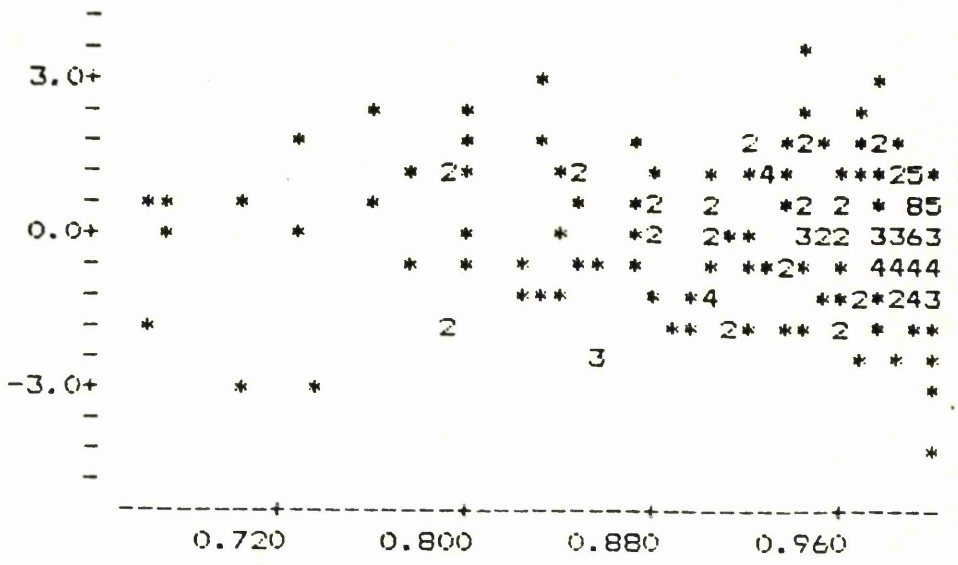
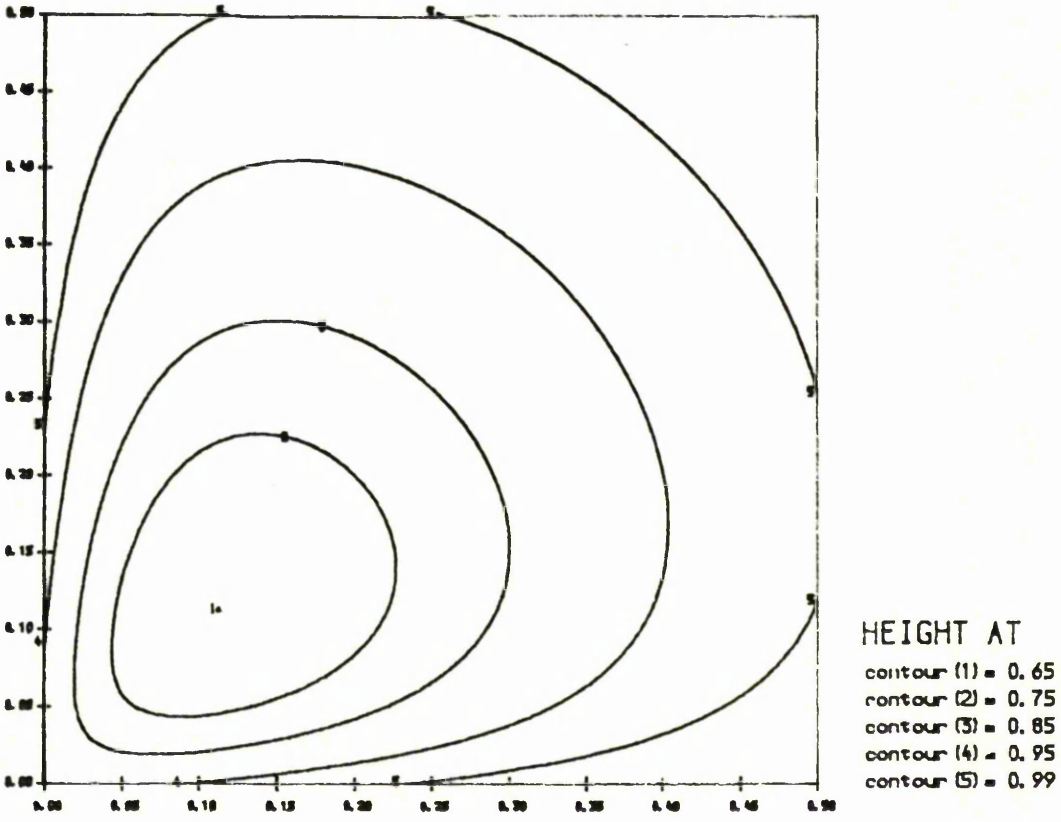


FIGURE (3.5) Estimated contour plot of P_4 vs θ_1 & θ_2



figure(3.2,m,s,l) shows the plotting of logit the data against Z_{2j} for all Z_{1i}; where

m=1 if Q_1 is the parameter to be modeled.

=2 if Q_2 is the parameter to be modeled.

s=1 if the true map function is Haldane.

=2 if the true map function is Eq(3).

l=1 if the assumed map function is Haldane.

=2 if the assumed map function is Kosambi.

=3 if the assumed map function is Eq(3).

3- Model M_2, a more general model fitting a different polynomial in Z_{2j} for each Z_{1i}. This model is mainly suggested by the plotting of logit(n_{ij}/m_{ij}) against Z_{2j} for each i. Some of these fourteen plots will be shown under the name of figure(3.3,m,s,l).

4- Model M_t, the true symmetric model.

(B)The result of the analysis will be presented in two tables and one figure, table(3.1,m,s,l) which shows the deviances and APA(s) of the different models , table(3.2,m,s,l) which shows the MLE of the parameters of the chosen model (always model M_1) and their corresponding confidence intervals, figure(3.4,m,s,l) which is a plot of Pearson residual against the fitted values of model M_1.

(C)But as we are really interested in providing estimates of P_1, P_W, P_4, then two possibilities will arise:

1- If we are modelling Q_2, which is equal to P_4, then three other tables and one further figure will be produced, table(3.3,P_4,s,l) which gives the individual APA(i) of model M_1, table(3.4,P_4,s,l) which is the same as the latter table but when restricted to the area of θ_1 and $\theta_2 \leqslant 0.3$, table(3.5,P_4,s,l) which shows how the overall APA within the restricted area is transmitted from the logit scale to the P_4 scale and finally

figure(3.5,P_4,s,1) which shows the estimated contour plot of P_4 against θ_1 and θ_2 according to model M_1.

2- If we are modelling Q_1, which is not of great interest on its own right but as a mean of estimating P_1 and P_W, three pairs of tables and one pair of figures will be produced with very similar notation and explanation as in (C)-1 but for P_1 and P_W, for example the first pair of tables would be table(3.3,P_1,s,1) and table(3.4,P_W,s,1) showing the individual APA(i) of the estimated P_1 and P_W respectively, and so on for the other tables and figures. Note that if $\tilde{n}_1(\theta_1,\theta_2)$ and $\tilde{n}_2(\theta_1,\theta_2)$ are the estimated linear predictors of Q_1 and Q_2 under model M_1 respectively, then by using formula(3.2) $\tilde{P}_1$ and $\tilde{P}_W$, the estimated P_1 and P_W respectively, would be

$$\tilde{P}_1 = \frac{\exp(\tilde{n}_1(\theta_1,\theta_2))}{[1+\exp(\tilde{n}_1(\theta_1,\theta_2))][1+\exp(\tilde{n}_2(\theta_1,\theta_2))]}$$

$$\tilde{P}_W = \frac{1}{[1+\exp(\tilde{n}_1(\theta_1,\theta_2))][1+\exp(\tilde{n}_2(\theta_1,\theta_2))]}$$

In that context the overall APA of $\tilde{P}_1$ (or $\tilde{P}_W$) would be redefined as follows

$$APA(\tilde{P}_1) = \left[\sum_{ij} [\text{logit}\left[\frac{r_{1ij}}{1000}\right] - \text{logit}(\tilde{P}_{1ij})]^2 / I*J \right]^{0.5}$$

Also notice, as we suspected that having this large number of figures and tables may be a burden on the reader, especially if situated within the text, we decided to put all of them at the end of the chapter.

Now as far as this sub-result-section is concerned, the two trial models M_1 and M_2 were defined as in (3.6) and (3.7) respectively. From Table(3.1,1,1,1) we can see that model M_1 produced a significant deviance and model M_2 did not. The APA of

M_1 which was equal to 0.54 was just 10% higher than that of M_2. Note that for this set of simulated data 28 observations were either equal to 0 or m_{ij}, they could be recognised from figure(3.1,1,1,1) as having the symbol A or H respectively. By using the chosen model, M_1, the overall APA of P_1 and P_W were equal to 0.36 and 0.45 respectively and would be reduced to 0.17 and 0.43 if calculated within the restricted area. From figures(3.5,P_1,1,1),(3.5,P_W,1,1) and (3.5) which are the estimated contour plots of P_1,P_W and P_4 against θ_1 and θ_2 respectively, we can see that the most informative area or more precisely the most rightly informative area of these two parameters is when both are between [0.05,0.20]. Within this range, P_4 is at its lowest level, which is between [0.65,0.75], P_1 is at its highest level, which is roughly between [0.25,0.35] and finally P_W is mainly at its lowest level which is roughly between [0.016,0.019]. Notice that, in all of those figures, and also the similar ones provided later, the height at the maximum or minimum of the function is not provided, although it could be roughly deduced from the contour's height key provided along side the plot. Most of these contour heights, have been chosen in regular step, the last contour height is the last one found within this stepwise search; this means, for example, that the maximum of figure(3.5,P1,1,1) is larger than 0.3 and less than 0.35.

(b)-Eq(3) is the true and Haldane is the assumed map function

(i)When Q_2 is modeled

The two trial models M_1 and M_2 were defined as in (3.6) and (3.7) respectively. Despite the substantial drop in the deviance from model M_0 to M_1 , M_1 still gave a significant result, M_2 on the other hand was insignificant. The significant result of model

M_1, shows itself as a pattern in the Pearson residual-fitted values plot in figure(3.4,2,3,1), which probably suggests that there is a missing term in the model. Nevertheless, from table(3.5,P_4,3,1) it seems that apart from the minimum value, the model gives a good fit to the data. Actually the overall APA of model M_1 was 0.35 which is 46% higher than the 0.24 of model M_2, but this is reduced to 0.19 if restricted to the interesting area of the θ(s).

(ii)<u>When Q_1 is modeled</u>

The first thing to notice is that, for this situation, there was a substantial number of observations which were equal to m_{ij}, which means that all the conclusive results were right. These observations are symbolised in figure(3.1,1,3,1) by the symbol H and they mainly correspond to the θ_1 or $\theta_2 \leqslant 0.05$. Faced with the difficult decision of either substituting them by smaller value, (eg) (m_{ij}-0.5), or, excluding them from the analysis, we decided on the later strategy. This decision was made in order to avoid a subjective choice of the substituted value, especially for such a wide scale of the observations. Actually no real need of a substitution is present here, as the data itself summarises the situation quite clearly suggesting that for θ_1 or $\theta_2 \leqslant 0.05$ if there is a conclusive decision then this decision is right. A 95% Bayesian interval for Q_{ij} using these observations could be provided as follows. If $\pi(Q)$ is the prior distribution of Q which for simplicity is assumed to be Uniform(0,1) for any combination(i,j), then the posterior density of Q_{ij} would be

$$p(Q_{1ij}|n_{ij},m_{ij}) \propto \pi(Q_1)\times L(Q_{1ij}|n_{ij},m_{ij})$$

$$\propto Q_{1ij}^{\,n_{ij}}\,(1-Q_{1ij})^{m_{ij}-n_{ij}}$$

which is the beta distribution with parameters (n_{ij}+1) and

$(m_{ij}-n_{ij}+1)$. But for these observations, $n_{ij}=m_{ij}$, meaning that the distribution will reach its maximum at $Q_{ij}=1$ and therefore a logical 95% Bayesian interval would be defined as follows

$$\int_{Q_{Lij}}^{1} p(Q_{ij}|n_{ij},m_{ij})\ dQ = 0.95$$

from which it is easy to see that the lower bound of the interval Q_{Lij}, will be equal to

$$Q_{Lij} = (0.05)^{(1/(m_{ij}+1))}$$

(i.e) the actual values of the interval will depend on m_{ij}, which in its turn depends on θ_1 and θ_2.Nevertheless fitting model M_1 and M_2 to the remaining observations produced an insignificant result for both models, as seen in table(3.1,1,3,1), (but rather unexpectedly model M_t was just significant).

From the estimated contour plots of P_1,P_W and P_4, we can see that the most rightly informative area is between [0.07,0.2], within this range P_4 was at its lowest level which was roughly between [0.57,0.65], P_1 was at its highest level which was roughly between [0.35,0.42]. As for P_W it was not at its lowest level but it is still less then 0.003, it is worth noticing as well that the shape of the estimated P_W seems to suggest that the true P_W is quite a flat function of θ_1 and θ_2 especially near its maximum. By comparing these three plots with their corresponding ones when Haldane was both the true and assumed map function, we can deduce that when the true map function is Eq(3), the amount of information that will be deduced from testing the gene orders will be slightly higher and that this information will be very unlikely to produce a wrong decision if we assume the Haldane map function. Whereas if the true map function was the Haldane the probability of wrong decision will be higher than its

corresponding one if the true map function was Eq(3).

(c)-Eq(3) is the true and Kosambi is the assumed map function

(i)When Q_2 is modeled

The trial model M_1, was slightly different than before, for this situation it was defined as follows

$$M_1: \quad \eta_{ij} = \alpha+\beta_1(Z_{1i}+Z_{2j})+\beta_2(Z_{1i}^2+Z_{2j}^2)+\beta_3(Z_{1i}*Z_{2j}) +\beta_4(Z_{1i}^3+Z_{2j}^3)+\beta_5(Z_{1i}^2*Z_{2j}+Z_{1i}*Z_{2j}^2)$$

(i.e) with the extra parameter β_5 which represents the quadratic interaction between Z_1 and Z_2. Model M_2 will still be defined as in (3.7) and still M_1 will be nested within M_2. From table(3.1,2,3,2) both models were significant which is again confirmed by the Pearson residual fitted value plot of M_1. From Table(3.5,P_4,3,2) it seems that model M_1 gave a reasonable approximation to the data.

(ii)When Q_1 is modeled

Model M_1 was slightly different than that of Q_2, the quadratic interaction term was not included although a fourth power term was, (i.e) the model was defined as follows

$$M_1: \eta_{ij} = \alpha+\beta_1(Z_{1i}+Z_{2j})+\beta_2(Z_{1i}^2+Z_{2j}^2)+\beta_3(Z_{1i}*Z_{2j})+\beta_4(Z_{1i}^3+Z_{2j}^3) +\beta_5(Z_{1i}^4+Z_{2j}^4) \qquad (3.8)$$

and therefore model M_2 had to be redefined as follows

$$M_2: \quad \eta_{ij} = \alpha_i+\beta_{1i}Z_{2j}+\beta_{2i}Z_{2j}^2+\beta_{3i}Z_{2j}^3+\beta_{4i}Z_{2j}^4 \qquad (3.9)$$
$$\text{for } i=1,2,\ldots,14$$

But as mentioned before only the significant terms would be included when fitting M_2 for each i and actually for this situation for all i β_{4i} was insignificant. From table(3.1,1,3,2) M_1 was significant whereas M_2 was not. The APA of P_1 was quite

acceptable as judged from table(3.5,P_1,3,2) although not the same could be said about P_W as judged from table(3.5,P_W,3,2).

From the estimated contour plots, the range of [0.07,0.17] still was the most informative range for both parameters θ_1 and θ_2. But if we compare these figures to their corresponding ones when Haldane was assumed and Eq(3) was still the true map function, it is clear that assuming kosambi will lead to a much lower inconclusive result, which is roughly between [0.37,0.45], a higher right decision, roughly with probability between [0.55,0.60], and also a higher wrong decision roughly with probability less then 0.008.

(d)-Eq(3) is the true and assumed map function

(i)When Q_2 is modeled

The trial model M_1 was defined as follows:

$$M_1: \quad \eta_{ij} = \alpha+\beta_1(Z_{1i}+Z_{2j})+\beta_2(Z_{1i}^2+Z_{2j}^2)+\beta_3(Z_{1i}^4+Z_{2j}^4)$$
$$+\beta_4(Z_{1i}^2*Z_{2j}^2)+\beta_5(Z_{1i}^3*Z_{2j}+Z_{1i}*Z_{2j}^3)$$

In this model no cubic term was included, although a fourth and all fourth interaction terms were. Model M_2 will be, then, defined as in (3.9). From table(3.1,2,3,3), both models gave significant results, although from table(3.5,P_4,3,3), it seems that model M_1 gave a reasonable approximation to the data.

(ii)When Q_1 is modeled

Model M_1 was defined as in (3.8), and therefore model M_2 will be defined as in (3.9). The latter model gave an insignificant result, as shown in table(3.1,1,3,3). The APA of P_1 was again quite acceptable although that of P_W was not quite, as judged form table(3.5,P_1,3,3) and table(3.5,P_W,3,3) respectively. From the estimated contour plots, the range of [0.07,0.17] was still the most informative range for both θ_1 and θ_2. A comparison

between these contours and the corresponding ones when the Haldane or the Kosambi map functions were assumed could be summarised as follows.

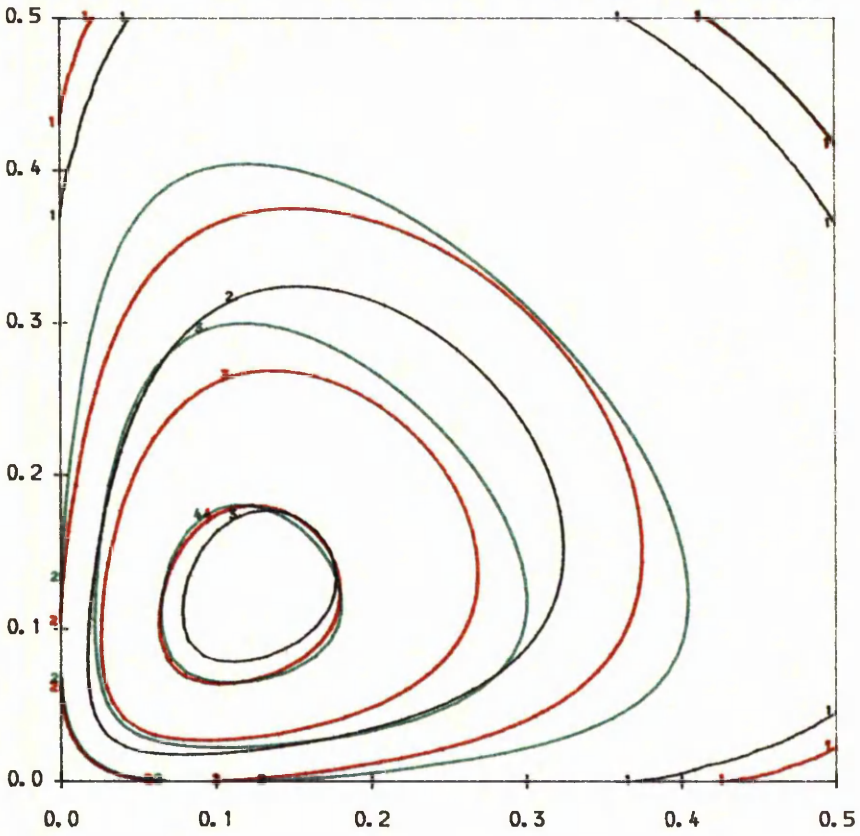
Summary

A contour plot for the estimated $P_1(\theta_1,\theta_2)$ and $P_W(\theta_1,\theta_2)$ when the true map function was Eq(3) and the assumed map function was either Haldane, Kosambi or Eq(3) is given in figure(3.6)(a) & (b) respectively, where each assumed map function's contours are drawn using different colours. From figure(3.6)(a), it seems quite clear that any of the first three contour heights; which were of height 0.001, 0.15 and 0.40 respectively, when the assumed map function is the Haldane, black curves, are contained within the corresponding contour heights when Kosambi is assumed, red curves, which in their turn are contained within the corresponding ones when Eq(3) is assumed, green curves. This means that as the assumed map function becomes closer and closer to the true one, more right decision will be made. The difference between the three assumed map functions becomes clearer as we approach the maximum of $\hat{P}_1$, which seems to lie at approximately $\theta_1=\theta_2=0.13$ for all map functions. Actually the nearest contours to the maximum in the plot –contour(3) for Haldane, contour(4) for Kosambi or Eq(3)– suggest that the square area of both θ_1 and θ_2 roughly in the range of [0.08,0.16] are at least as high as 0.40, 0.58 and 0.65 for the Haldane, Kosambi or Eq(3) respectively.

As for figure(3.6)(b), three different contour heights were drawn for each assumed map function. By looking first at the lower triangular half of the plot, we can see that the first contour, of height 0.001, 0.003 and 0.003 for the three assumed map functions, roughly coincided on top of each other; also the

FIGURE (3.6) A comparison between the estimated conclusive probabilities, for the different assumed map functions

(a) Estimated contour plot of P_1 vs θ_1 & θ_2



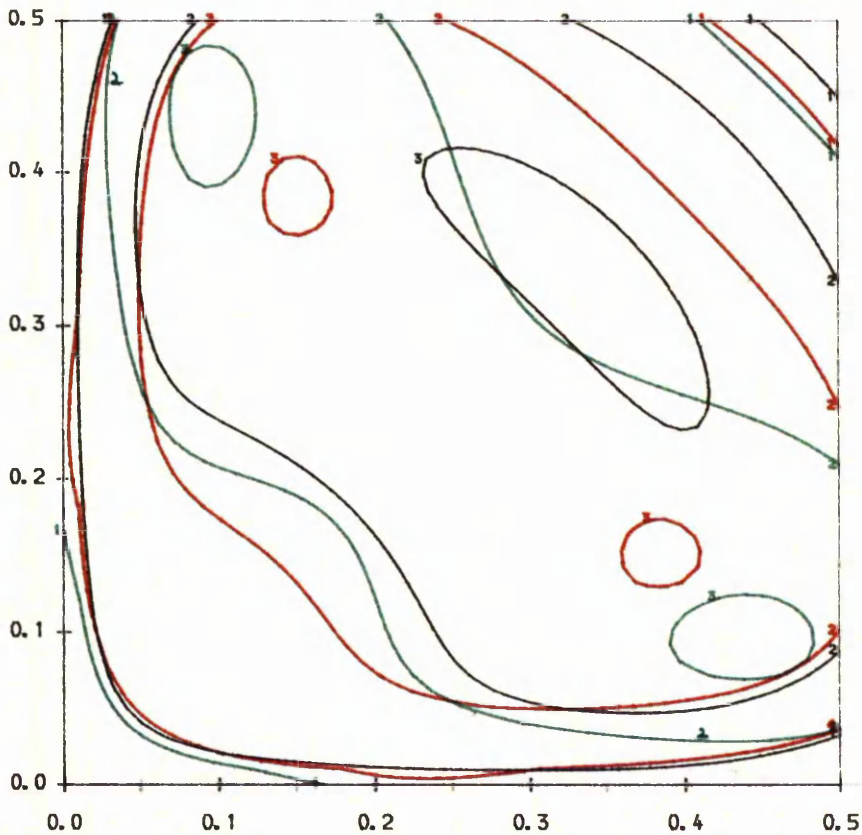
HALDANE
KOSAMBI
EQ (3)

HEIGHT AT
cont (1) = 0.001
cont (2) = 0.15
cont (3) = 0.40

the same &
cont (4) = 0.58

the same &
cont (4) = 0.65

(b) Estimated contour plot of P_w vs θ_1 & θ_2



HEIGHT AT
cont (1) = 0.001
cont (2) = 0.003
cont (3) = 0.0055

cont (1) = 0.003
cont (2) = 0.008
cont (3) = 0.012

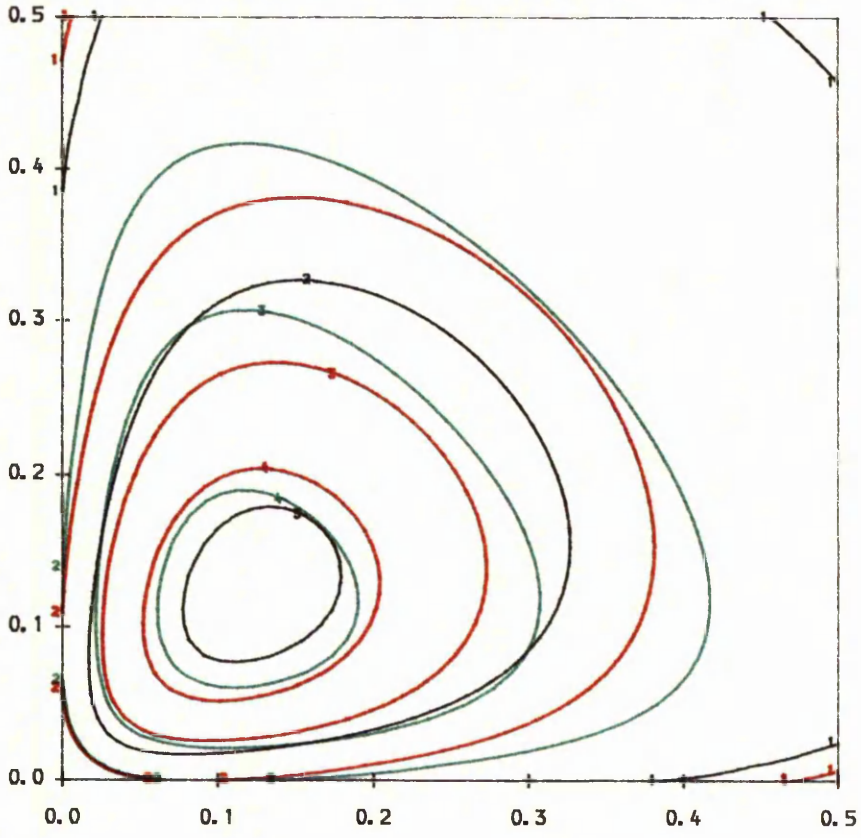
cont (1) = 0.003
cont (2) = 0.014
cont (3) = 0.023

same comment could be made about the second contour height 0.003, 0.008 and 0.014 respectively. This seems to suggest that as we approach toward the true map function, more and more wrong decisions are been made. This conclusion could be again reinforced by comparing the third contour height for each map function; where for the Haldane the area of θ_1 and θ_2 which is surrounded by this contour is at least as high as 0.0055, whereas the corresponding area for the Kosambi or Eq(3) are at least as high as 0.012 and 0.023 respectively. Another noticable feature of this figure, though not an important one because of the flatness of the estimated P_Ws, is the shift of the maximum area as we approach the true map function.

These comments could be summarised as follows, as the assumed map function becomes closer and closer to the true one, in general more and more conclusion will be made (notice the plot of $\hat{P}_4$, shown in figure(3.7)(a), which is a mirror image of the plot of $\hat{P}_1$). But although a big proportion of these conclusive results will be right conclusion, more and more wrong conclusion will be reached as well. A plot of the estimated probability of a right conclusion given a conclusive result, Q_1, for the three assumed map functions, shown in figure(3.7)(b), reinforce this last comment. The disappointment of this comment could perhaps be remedied by the quite small $\hat{P}_W$ for any of the assumed map functions. It is also worth emphasizing that, as the Haldane map function seems to be suggested in some of the recent genetical papers in order to be used in mapping multipoint loci data (Lathrop et al 1984, 85, 87), our result seems to suggest that by assuming this map function no real concern should be made about making a wrong decision, the only concern should be the small probability of producing any conclusive one.

FIGURE (3.7) A comparison between the estimated P_4 & Q_1 for the different assumed map functions

(a) Estimated contour plot of P_4 vs θ_1 & θ_2



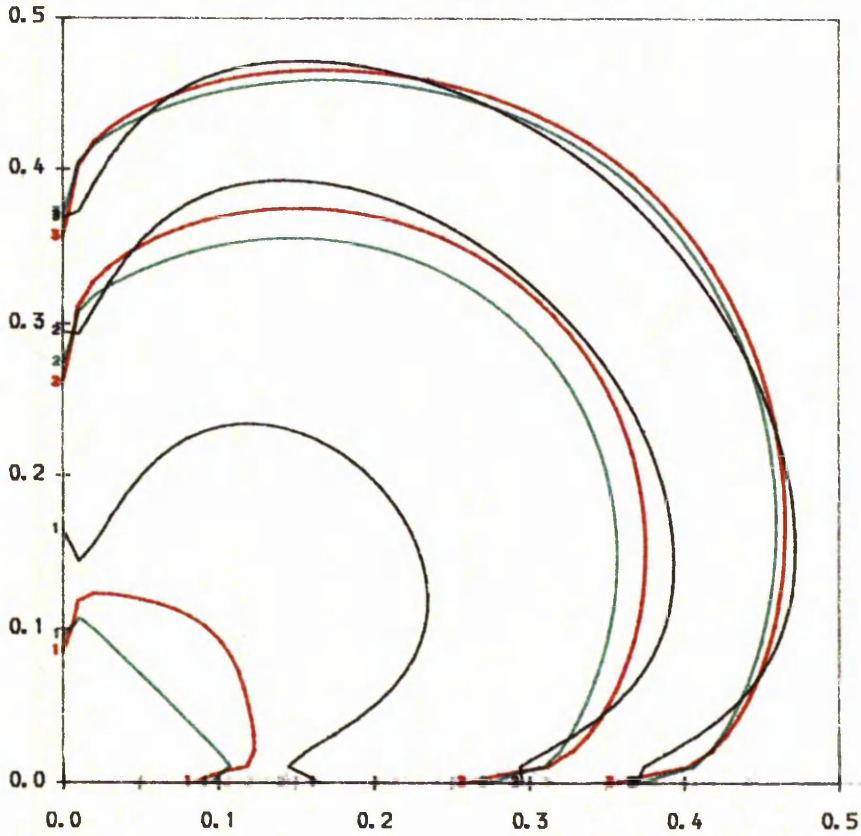
HALDANE
KOSAMBI
EQ (3)

HEIGHT AT
cont (1) = 0.999
cont (2) = 0.85
cont (3) = 0.60

the same &
cont (4) = 0.45

the same &
cont (4) = 0.35

(b) Estimated contour plot of Q_1 vs θ_1 & θ_2



HEIGHT AT
cont (1) = 0.99
cont (2) = 0.925
cont (3) = 0.75

HALDANE IS THE TRUE &

THE ASSUMED MAP FUNCTION

When Q_1 is modeled & Haldane is both the true & assumed map function

FIGURE (3.1,1,1,1) Discret contour plot of the data vs θ_1 & θ_2

```
0.50+  ***AD    D   C   C   C   B   C   *   A   *
   -   *HHAE    D   D   D   D   E   C   E   E   A
   -   H*DED    E   E   E   F   F   E   D   D   A
   -   HHFEF    F   F   E   F   F   E   F   D   C
   -   HEFEF    F   F   F   G   F   E   E   D   D
0.25+  GFFGF    G   F   G   G   F   G   E   D   D
   -   GFFGG    G   G   G   G   F   F   E   D   C
   -   FFFFF    G   G   G   G   G   G   E   D   C
   -   GFGGG    G   G   G   G   G   F   E   D   D

0.050+ GGGGG    G   G   F   G   F   E   F   F   E
   -   GGGGG    G   G   G   F   E   F   C   D   D
   -   GGGGG    F   G   G   E   G   F   E   E   A
   -   GGGGG    G   G   F   F   E   D   H   H   *
   -   GGGGG    G   F   F   H   H   H   A   *   *
0.000+
      +---------+---------+---------+---------+---------+---
     0.00      0.10      0.20      0.30      0.40      0.50
```

A = 0.0
B = 0.001---0.249
C = 0.250---0.499
D = 0.500---0.699
E = 0.700---0.799
F = 0.800---0.899
G = 0.900---0.999
H = 1.000
* = Q2 = 1.000

FIGURE (3.2,1,1,1) Logit of the data vs Z_{2j}

```
   -     *   *
3.5+     2   * 22*   * 2
   -     *   4 665   4 * 7 5
   -     4   2 2 3   6 8 2 564*
   -         * 224   * * 3 *4454
   -         * *2*   2 *   *2243
0.0+         *  2       * 2 2***22
   -                          *  4
   -                             *
     -------+---------+---------+--
          -4.0       -2.0       0.0
```

When Q_1 is modeled & Haldane is both the true & assumed map function

FIGURE (3.3,1,1,1) Logit of the data vs Z_{2j} for some Z_{1i}

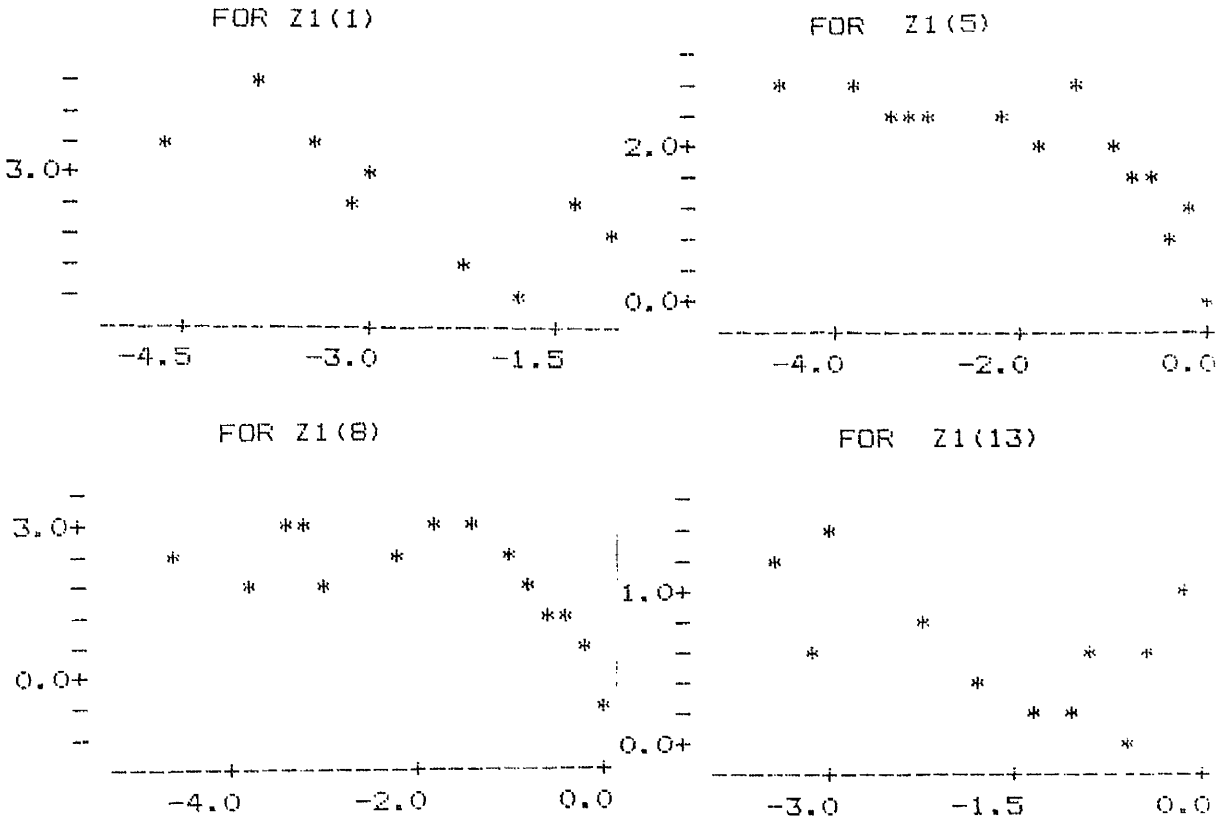
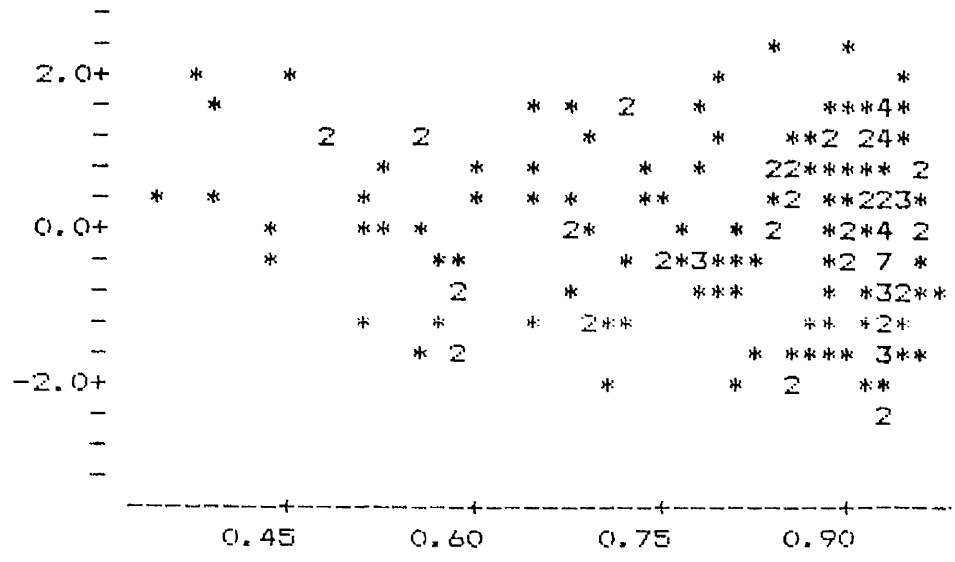


FIGURE (3.4,1,1,1) Pearson residuals vs the fitted values



When Q_1 is modeled and Haldane is both the true and assumed map function.

Table(3.1,1,1,1)A summary of the performance of different models

Model	Deviance	D.f	T.S	APA	$\frac{APA(M_1)}{APA(M_i)}$
M_0	763.5	167		1.16	
M_1	202.4	163	2.2	0.54	1.00
M_2	140.2	140	0.01	0.49	1.10
M_t	58.0	75	-1.4	0.33	1.64

Table(3.2,1,1,1)MLE & 95% I.E for the parameters of Model M_1

Parameters	MLE	95% Interval estimates
α	-1.544	-1.935 ; -1.153
β_1	-2.785	-3.124 ; -2.446
β_2	-1.228	-1.395 ; -1.061
β_3	0.229	0.174 ; 0.284
β_4	-0.139	-0.164 ; -0.114

Table(3.3,P_1,1,1) Individual APA(M_1,i)

i=1,2,...7	0.32	0.20	0.28	0.16	0.27	0.23	0.22
i=8,9,..14	0.16	0.24	0.45	0.50	0.42	0.78	0.57

(The overall APA is equal 0.36).

Table(3.4,P_1,1,1) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.32	0.20	0.18	0.14	0.12	0.13	0.15
i=8,9,10	0.14	0.12	0.13				

(The overall APA is equal 0.17).

When Q_1 is modeled and Haldane is both the true and assumed map function.

Table(3.5,P_1,1,1)Examples showing the errors made by model M_1

$\mu^*=\frac{n_1}{1000}$	Logit(μ^*)	η^*_l= Logit(μ^*) -0.17	η^*_u= Logit(μ^*) +0.17	$\frac{Exp(\eta^*_l)}{1+Exp(\eta^*_l)}$	$\frac{Exp(\eta^*_u)}{1+Exp(\eta^*_u)}$
0.002	-6.213	-6.383	-6.043	0.002	0.002
0.049	-2.966	-3.136	-2.796	0.042	0.058
0.102	-2.175	-2.345	-2.005	0.087	0.119
0.162	-1.643	-1.813	-1.473	0.140	0.186
0.348	-0.628	-0.798	-0.458	0.310	0.387

Table(3.3,P_W,1,1) Individual APA(M_1,i)

i=1,2,...7	0.47	0.50	0.46	0.43	0.28	0.33	0.38
i=8,9,..14	0.39	0.43	0.57	0.52	0.60	0.60	0.36

(The overall APA is equal 0.45).

Table(3.4,P_W,1,1) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.47	0.50	0.46	0.42	0.25	0.30	0.41
i=8,9,10	0.38	0.44	0.62				

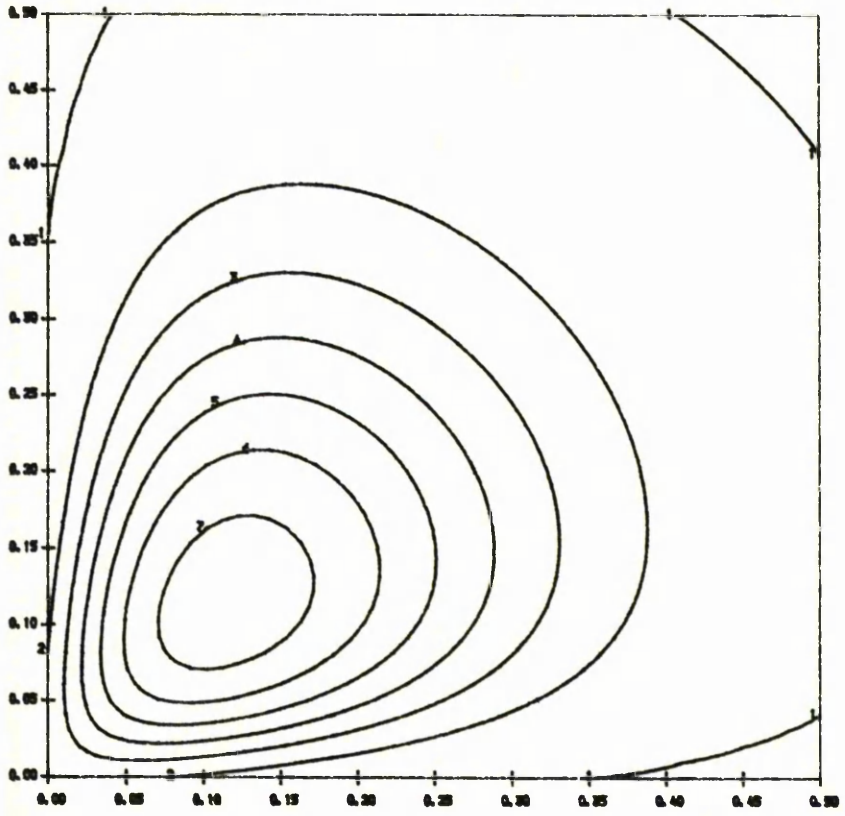
(The overall APA is equal 0.43).

Table(3.5,P_W,1,1)Examples showing the errors made by model M_1

$\mu^*=\frac{n_W}{1000}$	Logit(μ^*)	η^*_l= Logit(μ^*) -0.43	η^*_u= Logit(μ^*) +0.43	$\frac{Exp(\eta^*_l)}{1+Exp(\eta^*_l)}$	$\frac{Exp(\eta^*_u)}{1+Exp(\eta^*_u)}$
0.001	-6.907	-7.337	-6.477	0.001	0.002
0.004	-5.517	-5.947	-5.087	0.003	0.006
0.009	-4.701	-5.131	-4.271	0.006	0.014
0.014	-4.255	-4.685	-3.825	0.009	0.021
0.025	-3.664	-4.094	-3.234	0.016	0.038

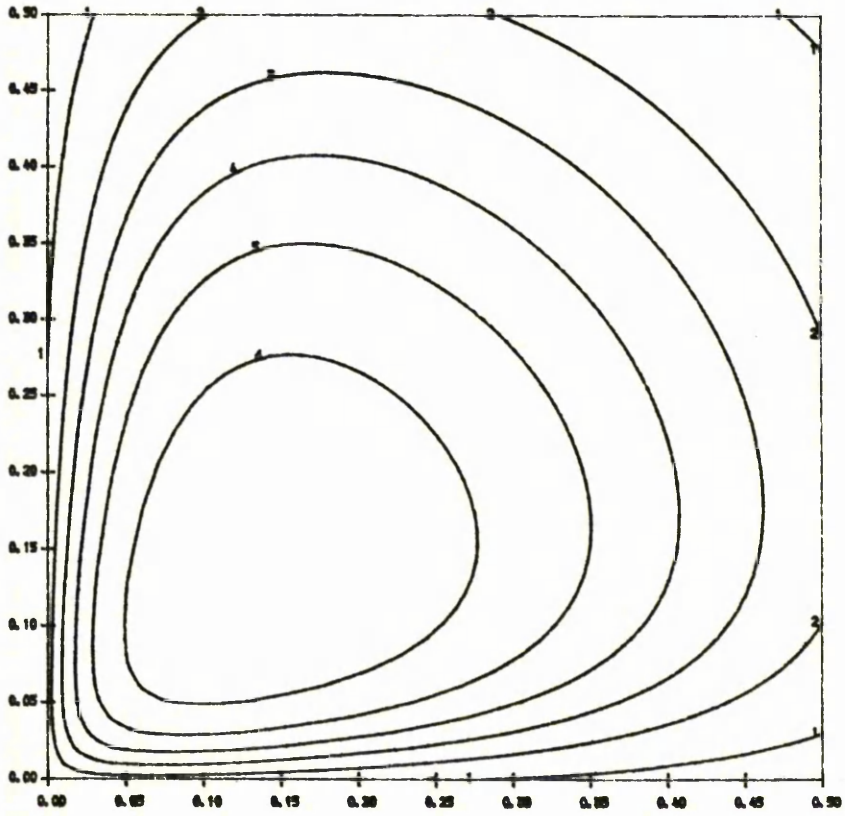
When Q_1 is modeled & Haldane is both the true & assumed map function

FIGURE (3.5, P1, 1, 1) Estimated contour plot of P_1 vs θ_1 & θ_2



HEIGHT AT
contour (1) = 0.001
contour (2) = 0.05
contour (3) = 0.1
contour (4) = 0.15
contour (5) = 0.2
contour (6) = 0.25
contour (7) = 0.3

FIGURE (3.5, PW, 1, 1) Estimated contour plot of P_w vs θ_1 & θ_2



HEIGHT AT
contour (1) = 0.001
contour (2) = 0.004
contour (3) = 0.007
contour (4) = 0.010
contour (5) = 0.013
contour (6) = 0.016

EQ(3) IS THE TRUE &

HALDANE IS THE ASSUMED MAP FUNCTION

When Q_2 is modeled & Eq(3) is the true & Haldane is the assumed map function

FIGURE (3.1,2,3,1) Discret contour plot of the data vs θ_1 & θ_2

```
0.50+  ***AA  A  A  A  A  A  A  A  *  *
   -   AAAAA  A  A  A  A  A  A  A  A  *
   -   AAAAA  A  A  A  A  A  A  A  A  A
   -   AAAAA  A  B  B  A  A  A  A  A  A
   -   AAAAA  B  B  B  B  A  A  A  A  A
0.25+  AAAAB  C  D  D  C  B  A  A  A  A
   -   AAABB  D  E  D  C  B  B  A  A  A
   -   AABBC  E  E  E  D  B  B  A  A  A
   -   ABBBC  D  E  D  C  B  A  A  A  A

0.050+ ABBBC  C  C  B  B  A  A  A  A  A
   -   ABBBB  C  B  B  A  A  A  A  A  A
   -   ABBBB  B  B  A  A  A  A  A  A  A
   -   AAABB  B  A  A  A  A  A  A  A  *
   -   AAAAA  A  A  A  A  A  A  A  A  *
0.000+
      +--------+--------+--------+--------+--------+
     0.00     0.10     0.20     0.30     0.40     0.50
```

* = 02=1.000
A = 0.900---0.999
B = 0.800---0.899
C = 0.700---799
D = 0.800---0.899
E = 0.500---0.699

FIGURE (3.2,2,3,1) Logit of the data vs Z_{2j}

```
   -
   -                                    ***
   -      3   2 22*                    **232
 5.0+     *   * * *                    **3242
   -      4   2 *2*    2 2 3 323554
   -      5   3 323    3 2 2 2444
   -          5 775    3 3 3 443
   -             *3    5 4 5 42
 0.0+                  * 3 *
      ------+--------+--------+--
           -4.0     -2.0      0.0
```

FIGURE (3.3,2,3,1) Logit of the data vs Z_{2j} for some Z_{1i}

```
          FOR Z1(2)                          FOR Z1(7)
   -                                  -
 6.0+                               3.0+
   -                                  -
   -                                  -
   -                                  -
 3.0+                               0.0+
   -                                  -
   -                                  -
   -                                  -
   +--------+--------+--------+      +--------+--------+--------+
  -4.0      -2.0      0.0            -4.0      -2.0      0.0
```

When Q_2 is modeled and Eq(3) is the true and Haldane is the assumed map function.

Table(3.1,2,3,1)A summary of the performance of different models

Model	Deviance	D.f	T.S	APA	$\frac{APA(M_1)}{APA(M_i)}$
M_0	21820.0	187		1.83	
M_1	366.7	183	9.6	0.35	1.00
M_2	160.0	134	1.6	0.24	1.46
M_t	69.1	87	-1.36	0.17	2.06

Table(3.2,2,3,1)MLE & 95% I.E for the parameters of Model M_1

Parameters	MLE	95% Interval estimates
α	7.620	7.469 ; 7.771
β_1	4.536	4.417 ; 4.655
β_2	1.905	1.851 ; 1.959
β_3	-0.391	-0.409 ; -0.373
β_4	0.181	0.174 ; 0.188

Table(3.3,P_4,3,1) Individual APA(M_1,i)

i=1,2,...7	0.53	0.33	0.25	0.26	0.23	0.15	0.18
i=8,9,..14	0.15	0.23	0.26	0.17	0.37	0.74	0.52

(The overall APA is equal 0.35).

Table(3.4,P_4,3,1) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.38	0.20	0.17	0.17	0.10	0.16	0.19
i=8,9,10	0.12	0.15	0.13				

(The overall APA is equal 0.19).

Table(3.5,P_4,3,1)Examples showing the errors made by model M_1

$\mu^*=\frac{n_4}{1000}$	Logit(μ^*)	n^*_l= Logit(μ^*) -0.19	n^*_u= Logit(μ^*) +0.19	$\frac{Exp(n^*_l)}{1+Exp(n^*_l)}$	$\frac{Exp(n^*_u)}{1+Exp(n^*_u)}$
0.508	0.032	-0.158	0.222	0.461	0.555
0.805	1.418	1.228	1.608	0.773	0.833
0.881	2.002	1.812	2.192	0.860	0.900
0.941	2.769	2.579	2.959	0.930	0.951
0.997	5.806	5.616	5.996	0.996	0.998

When Q_2 is modeled & Eq (3) is the true & Haldane is the assumed map function

FIGURE (3.4,2,3,1) Pearson residuals vs the fitted values

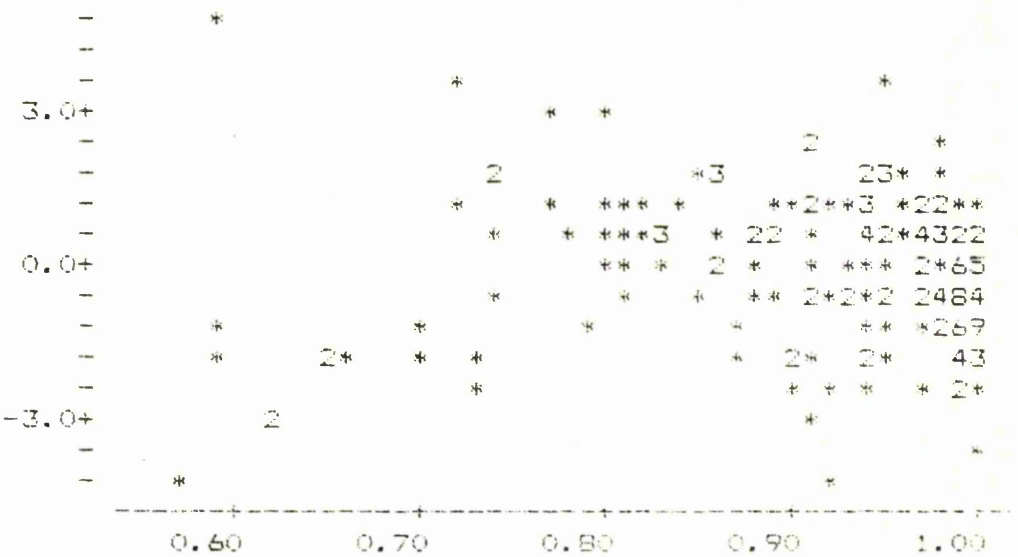
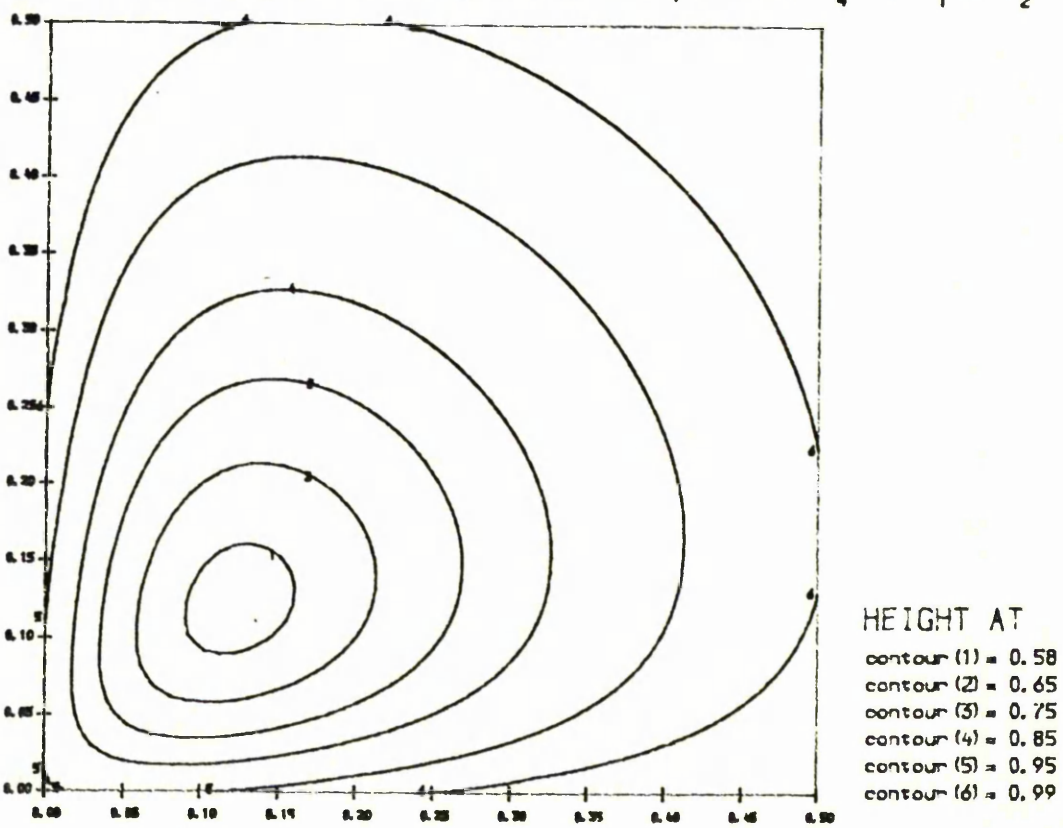


FIGURE (3.5,2,5,1) Estimated contour plot of P_4 vs θ_1 & θ_2



When Q_1 is modeled & Eq(3) is the true & Haldane is the assumed map function

FIGURE (3.1,1,3,1) Discret contour plot of the data vs θ_1 & θ_2

```
0.50+  ***DA    C    C    C    D    C    D    E    *    *
   -   HHDCH    E    E    E    E    D    E    C    D    *
   -   AHFFH    G    F    G    F    F    D    E    D    H
   -   HHGHG    G    G    G    G    F    F    G    D    C
   -   DHGGG    G    G    H    G    G    G    H    E    D
0.25+  HHGGG    G    G    G    G    G    G    E    E    F
   -   HGHGG    G    G    G    G    G    G    G    F    C
   -   HHHGH    G    G    G    G    G    G    F    E    D
   -   HHGHG    G    H    G    G    G    G    E    D    E

0.050+ HHGHG    H    G    G    G    G    G    G    F    D
   -   HHHHH    G    H    H    G    G    H    G    D    C
   -   HHHGH    H    H    G    H    G    H    F    H    D
   -   HHHHH    G    H    H    G    H    H    H    H    *
   -   HHHHH    H    H    H    G    H    H    H    H    *
0.000+
      +---------+---------+---------+---------+---------+
     0.00      0.10      0.20      0.30      0.40      0.50
```

A = 0.0
B = 0.001---0.249
C = 0.250---0.499
D = 0.500---0.699
E = 0.700---0.799
F = 0.800---0.899
G = 0.900---0.999
H = 1.000
* = 02 = 1.000

FIGURE (3.2,1,3,1) Logit of the data vs Z_{2J}

```
   -                    2 * *
5.0+              **5   2 5 3 32
   -           * **     2    5 5***
   -    *      *   2    * *    *44*
   -              *22       * 2 **45
   -              *      2 *     22*37*
0.0+              ***    * *      ** 24
   -               *            *     *3
   -
     ------+---------+---------+
          -4.0      -2.0       0.0
```

When Q_1 is modeled & Eq(3) is the true & Haldane is the assumed map function

FIGURE (3.3,1,3,1) Logit of the data vs Z_{2j} for some Z_{1r}

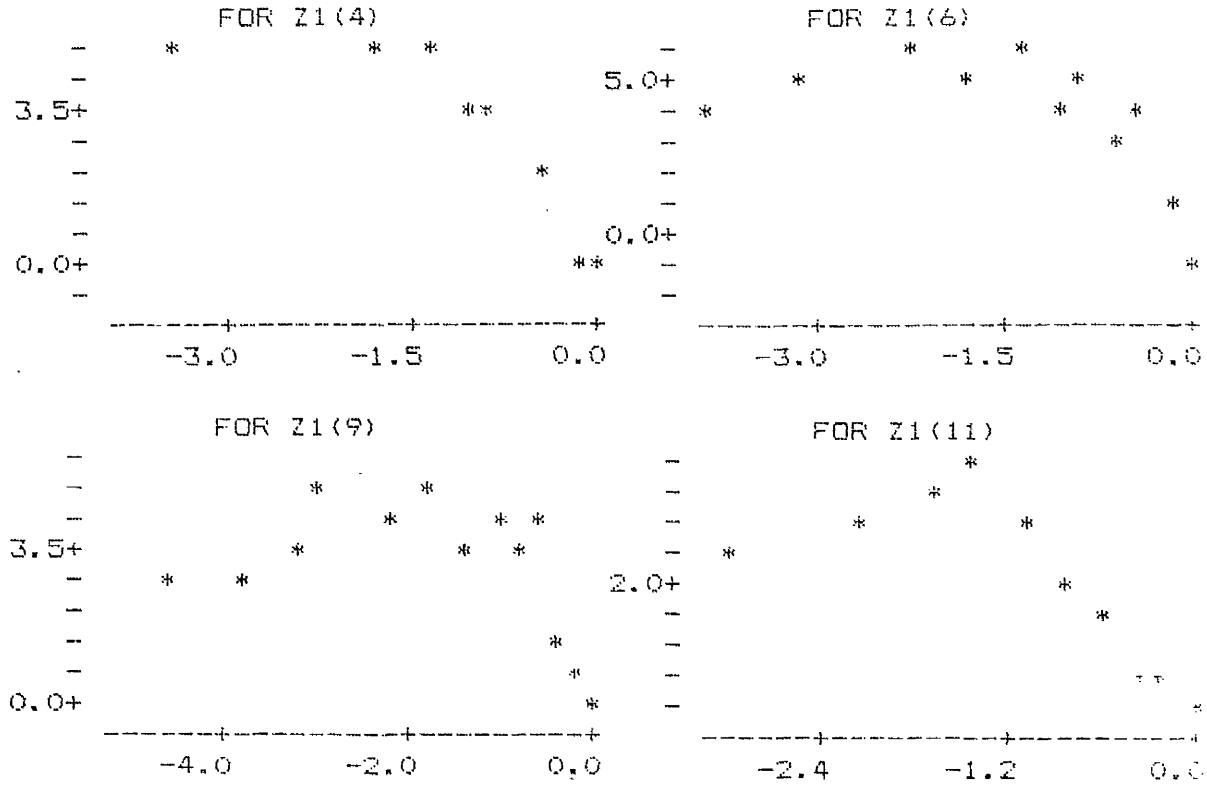
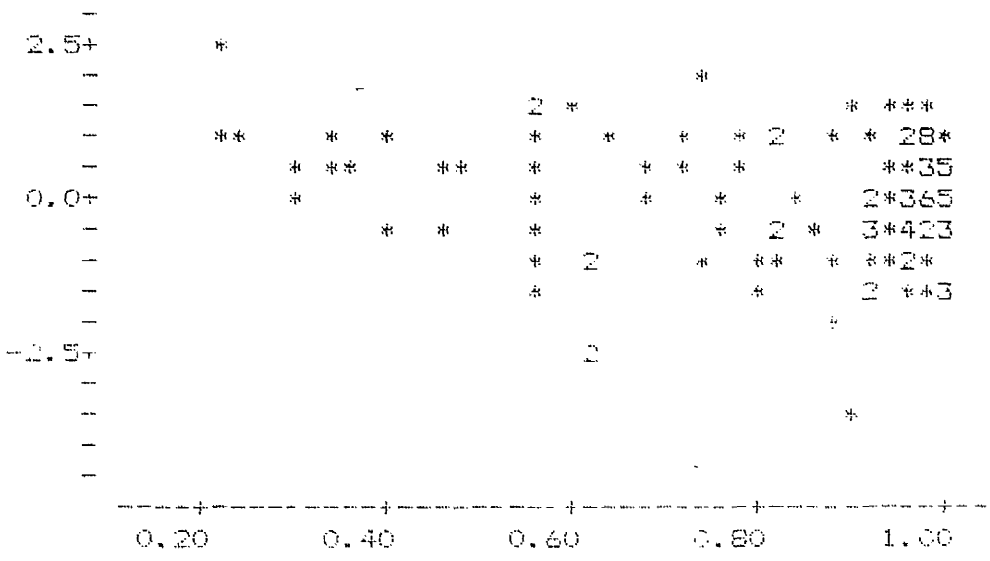


FIGURE (3.4,1,3,1) Pearson residuals vs the fitted values



When Q_1 is modeled and Eq(3) is the true and Haldane is the assumed map function.

Table(3.1,1,3,1)A summary of the performance of different models

Model	Deviance	D.f	T.S	APA	$\frac{APA(M_1)}{APA(M_i)}$
M_0	957.7	117		2.05	
M_1	116.1	113	0.2	0.72	1.00
M_2	96.7	93	0.3	0.64	1.10
M_t	60.1	42	1.97	0.53	1.36

Table(3.2,1,3,1)MLE & 95% I.E for the parameters of Model M_1

Parameters	MLE	95% Interval estimates
α	-2.923	-3.530 ; -2.316
β_1	-4.801	-5.492 ; -4.110
β_2	-2.008	-2.421 ; -1.595
β_3	0.425	0.265 ; 0.585
β_4	-0.220	-0.296 ; -0.144

Table(3.3,P_1,3,1) Individual APA(M_1,i)

i=1,2,...7	0.80	0.16	0.43	0.41	0.08	0.14	0.16
i=8,9,..14	0.24	0.22	0.25	0.35	0.64	0.52	0.86

(The overall APA is equal 0.42).

Table(3.4,P_1,3,1) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.80	0.16	0.18	0.15	0.09	0.13	0.16
i=8,9,10	0.14	0.15	0.11				

(The overall APA is equal 0.18).

When Q_1 is modeled and Eq(3) is the true and Haldane is the assumed map function.

Table(3.5,P_1,3,1)Examples showing the errors made by model M_1

$\mu^*=\frac{n_1}{1000}$	Logit(μ^*)	n^*_l= Logit(μ^*) -0.18	n^*_u= Logit(μ^*) +0.18	$\frac{Exp(n^*_l)}{1+Exp(n^*_l)}$	$\frac{Exp(n^*_u)}{1+Exp(n^*_u)}$
0.002	-6.213	-6.393	-6.033	0.002	0.002
0.058	-2.788	-2.968	-2.608	0.049	0.069
0.119	-2.002	-2.182	-1.822	0.101	0.139
0.194	-1.424	-1.604	-1.244	0.167	0.224
0.490	-0.040	-0.220	0.140	0.445	0.535

Table(3.3,P_W,3,1) Individual APA(M_1,i)

i=1,2,...7	1.01	0.08	0.42	0.45	0.54	0.81	0.42
i=8,9,..14	0.49	0.69	0.56	0.22	0.60	0.56	0.67

(The overall APA is equal 0.57).

Table(3.4,P_W,3,1) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leqslant 0.3$.

i=1,2,...7	1.01	0.08	0.43	0.42	0.59	0.81	0.48
i=8,9,10	0.49	0.57	0.53				

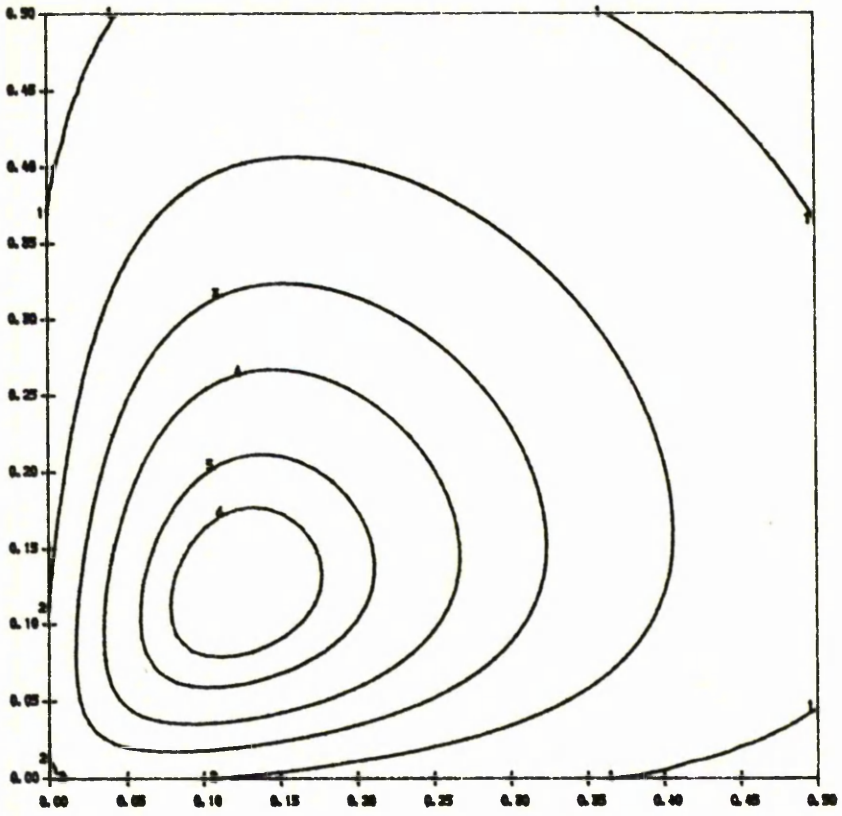
(The overall APA is equal 0.57).

Table(3.5,P_W,3,1)Examples showing the errors made by model M_1

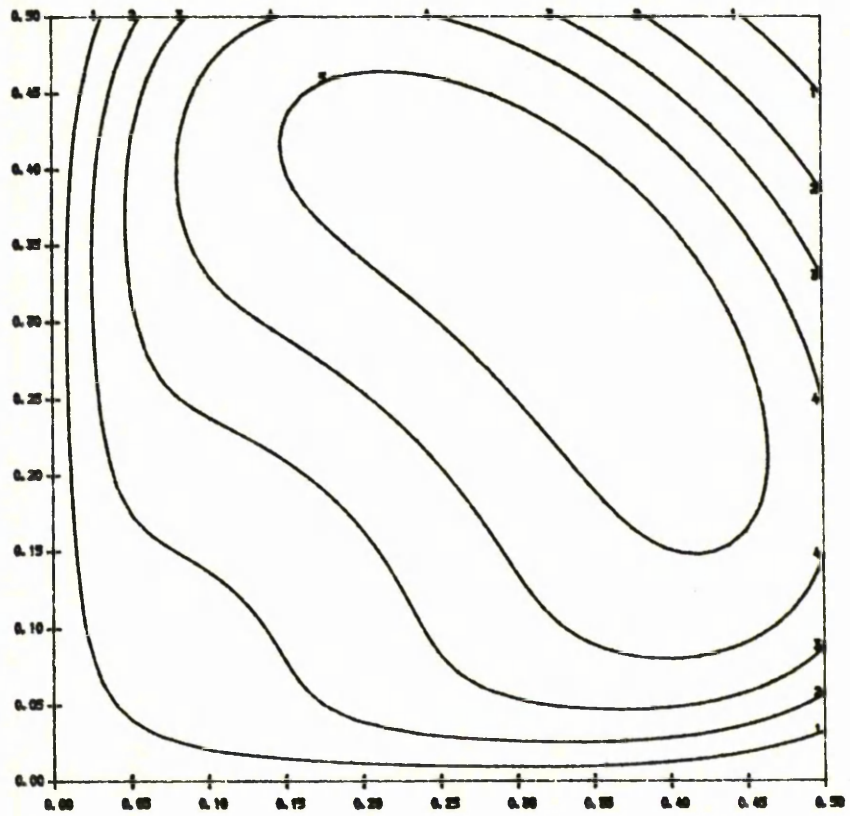
$\mu^*=\frac{n_W}{1000}$	Logit(μ^*)	n^*_l= Logit(μ^*) -0.57	n^*_u= Logit(μ^*) +0.57	$\frac{Exp(n^*_l)}{1+Exp(n^*_l)}$	$\frac{Exp(n^*_u)}{1+Exp(n^*_u)}$
0.001	-6.907	-7.477	-6.337	0.001	0.002
0.002	-6.213	-6.783	-5.643	0.001	0.004
0.007	-4.955	-5.525	-4.385	0.004	0.012

When Q_1 is modeled & Eq(3) is the true & Haldane is the assumed map function

FIGURE(3.5,P1,3,1) Estimated contour plot of P_1 vs θ_1 & θ_2



FIGURE(3.5,PW,3,1) Estimated contour plot of P_w vs θ_1 & θ_2



EQ(3) IS THE TRUE &

KOSAMBI IS THE ASSUMED MAP FUNCTION

When Q_2 is modeled & Eq (3) is the true & Kosambi is the assumed map function

FIGURE (3.1,2,3,2) Discret contour plot of the data vs θ_1 & θ_2

```
 0.50+   *AAAA     A      A      A      A      A      A      A      A      *
    -    AAAAA     A      A      A      A      A      A      A      A      A
    -    AAAAA     A      B      B      A      A      A      A      A      A
    -    AAAAA     B      C      B      B      *      A      A      A      A
    -    AAABB     C      D      D      C      B      B      A      A      A
 0.25+   AABBC     E      E      E      D      C      B      A      A      A
    -    ABCDD     F      F      F      E      C      B      B      A      A
    -    BCCEE     F      F      F      E      D      C      B      A      A
    -    BCDEE     F      F      F      E      C      B      A      A      A

0.050+   BBCDE     E      E      D      C      B      A      A      A      A
    -    ABCCD     E      D      C      B      B      A      A      A      A
    -    ABBCC     D      C      C      B      A      A      A      A      A
    -    ABBBC     C      C      B      B      A      A      A      A      A
    -    AAABA     B      B      A      A      A      A      A      A      *
0.000+
       +---------+---------+---------+---------+---------+--
      0.00      0.10      0.20      0.30      0.40      0.50

       * = Q2=1.000
       A = 0.900---0.999          D = 0.600---0.699
       B = 0.800---0.899          E = 0.450---0.599
       C = 0.700---0.799          F = 0.300---0.449
```

FIGURE (3.2,2,3,2) Logit of the data vs Z_{2J}

```
     -          *                   *
     -                            *22
  5.0+    3     *  *  *            233
     -    *     * 22*    *      *23356
     -    4     3 *2*    2  2  3 23544
     -    5     4 433    2  2  2 4544
     -          4 665    3  4  5 442
  0.0+             *3    4  4  4 3
     -                   2  2
        ------+---------+---------+-
            -4.0      -2.0       0.0
```

FIGURE (3.3,2,3,2) Logit of the data vs Z_{2J} for some Z_{11}

```
                FOR Z1(3)                               FOR Z1(10)
     -                                  *   -
     -                                 *    -         *
     -                                      -              *                *
 3.5+                                 *  2.5+                *
     -      *                        *      -                  **            *
     -                               *      -                               *
     -          *  **            *  *       -                       *    *  *
     -              *     *  *              -                         *
  0.0+                                   0.0+
       ------+---------+---------+              ------+---------+-----------
           -4.0      -2.0       0.0                 -4.0      -2.0       0.(
```

When Q_2 is modeled and Eq(3) is the true and Kosambi is the assumed map function.

Table(3.1,2,3,2)A summary of the performance of different models

Model	Deviance	D.f	T.S	APA	$\frac{APA(M_1)}{APA(M_i)}$
M_o	37950.0	192		1.77	
M_1	403.8	187	11.2	0.25	1.00
M_2	209.4	138	4.3	0.19	1.32
M_t	71.0	90	-1.4	0.12	2.12

Table(3.2,2,3,2)MLE & 95% I.E for the parameters of Model M_1

Parameters	MLE	95% Interval estimates
α	6.726	6.584 ; 6.868
β_1	4.257	4.142 ; 4.372
β_2	1.771	1.729 ; 1.813
β_3	-0.538	-0.599 ; -0.477
β_4	0.175	0.169 ; 0.181
β_5	-0.032	-0.038 ; -0.026

Table(3.3,P_4,3,2) Individual APA(M_1,i)

i=1,2,...7	0.16	0.26	0.20	0.17	0.15	0.14	0.14
i=8,9,..14	0.14	0.15	0.21	0.22	0.26	0.45	0.56

(The overall APA is equal 0.25).

Table(3.4,P_4,3,2) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.14	0.16	0.09	0.09	0.12	0.12	0.15
i=8,9,10	0.12	0.12	0.13				

(The overall APA is equal 0.11).

Table(3.5,P_4,3,2)Examples showing the errors made by model M_1

$\mu^*=\frac{n_4}{1000}$	Logit(μ^*)	n^*_l= Logit(μ^*) -0.11	n^*_u= Logit(μ^*) +0.11	$\frac{Exp(n^*_l)}{1+Exp(n^*_l)}$	$\frac{Exp(n^*_u)}{1+Exp(n^*_u)}$
0.307	-0.814	-0.924	-0.704	0.284	0.331
0.630	0.532	0.422	0.642	0.604	0.655
0.758	1.142	1.032	1.252	0.737	0.778
0.881	2.002	1.892	2.112	0.869	0.892
0.974	3.623	3.513	3.733	0.971	0.977

When Q_2 is modeled & Eq(3) is the true & Kosambi is the assumed map function

FIGURE (3.4,2,3,2) Pearson residuals vs the fitted values

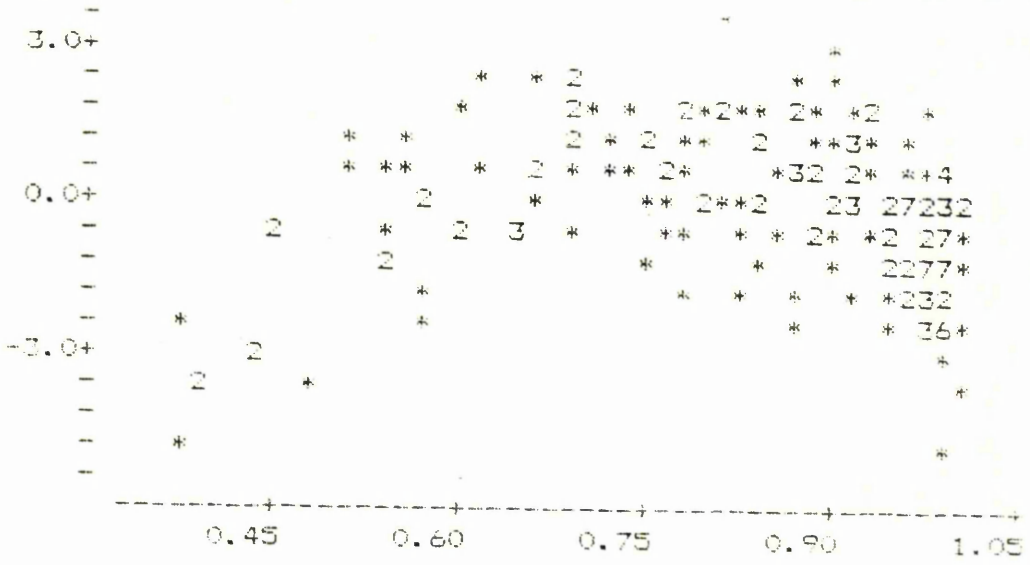
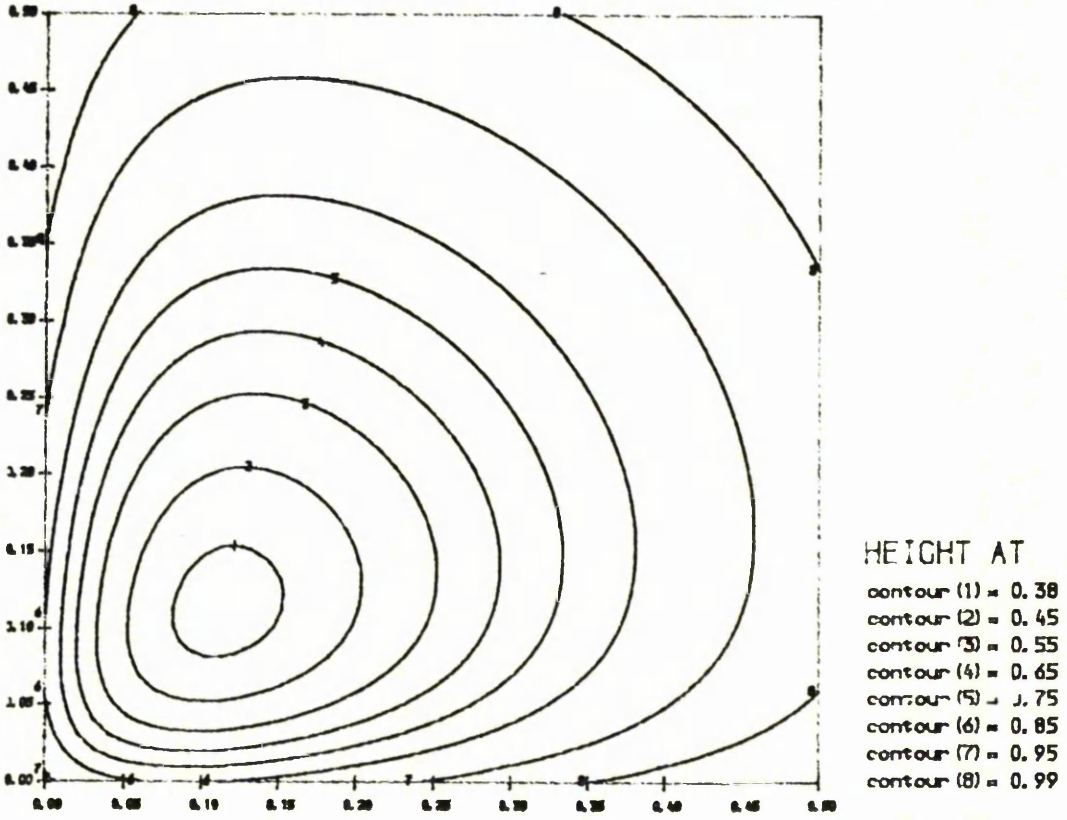


FIGURE (3.5,2,3,1) Estimated contour plot of P_4 vs θ_1 & θ_2



When Q_1 is modeled & Eq(3) is the true & Kosambi is the assumed map function

FIGURE (3.1,1,3,2) Discret contour plot of the data vs θ_1 & θ_2

```
0.50+   *AEDC    D    D    D    C    D    D    E    C    *
   -    HEDDE    D    E    E    D    D    E    C    C    H
   -    EFFGG    F    F    F    F    F    E    E    D    D
   -    FFGGG    G    G    G    G    G    F    G    D    D
   -    FGGGG    G    G    G    G    G    F    G    E    C
0.25+   GGGGG    G    G    G    G    G    G    F    D    C
   -    HGGGG    G    G    G    G    G    G    F    E    C
   -    GGGGG    G    G    G    G    G    G    F    E    D
   -    GHGGG    G    G    G    G    G    G    F    E    D

0.050+  HHGHG    H    G    G    G    G    F    F    E    D
   -    HHGHH    G    G    G    G    G    F    F    G    D
   -    HHHGH    G    G    G    G    G    G    E    F    D
   -    HHHHH    G    G    G    G    F    F    F    H    H
   -    HHHHH    H    H    G    G    G    D    E    H    *
0.000+
        +---------+---------+---------+---------+---------+--
      0.00      0.10      0.20      0.30      0.40      0.50
```

A = 0.0
B = 0.001---0.249
C = 0.250---0.499
D = 0.500---0.699
E = 0.700---0.799
F = 0.800---0.899
G = 0.900---0.999
H = 1.000
* = 02 = 1.000

FIGURE (3.2,1,3,2) Logit of the data vs Z_{2j}

```
   -
 6.0+          2*2   2
   -        3 2**   5 5 3 22
   -    *     232   2 5 7 62*
   -    2   2 *43   * 2   3796
   -    *   2 2 *   3 * 2 223682
 0.0+   *     ***     * * ***237
   -                      *2
   -
      ------+---------+---------+-
          -4.0      -2.0       0.0
```

When Q_1 is modeled & Eq(3) is the true & Kosambi is the assumed map function

FIGURE (3.3,1,3,2) Logit of the data vs Z_{2j} for some Z_{1i}

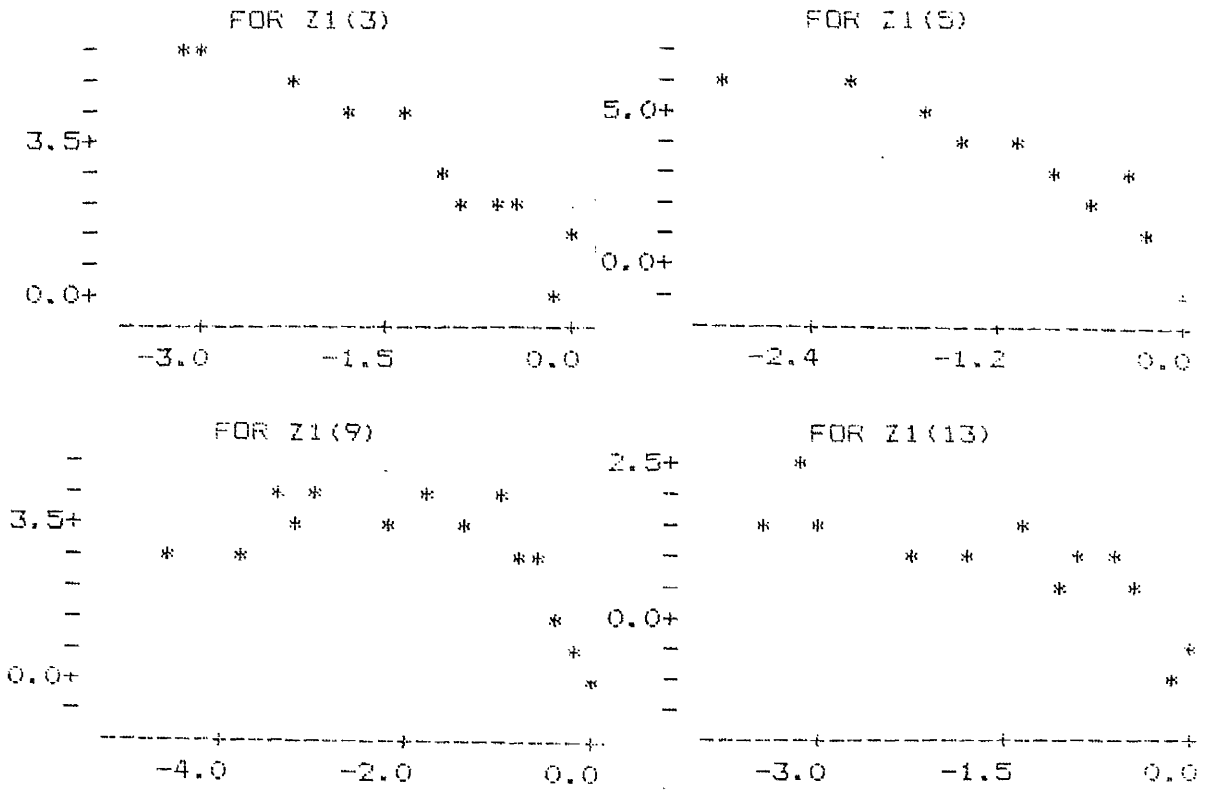
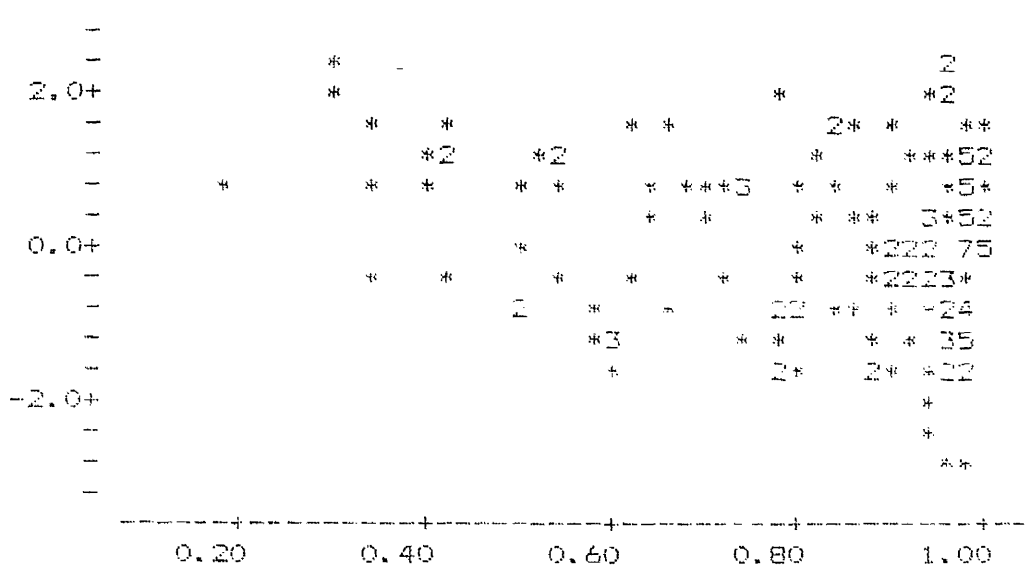


FIGURE (3.4,1,3,2) Pearson residuals vs the fitted values



When Q_1 is modeled and Eq(3) is the true and Kosambi is the assumed map function.

Table(3.1,1,3,2)A summary of the performance of different models

Model	Deviance	D.f	T.S	APA	$\frac{APA(M_1)}{APA(M_i)}$
M_0	1986.0	160		1.90	
M_1	222.5	155	3.8	0.60	1.00
M_2	145.2	130	0.9	0.48	1.25
M_t	84.0	72	1.0	0.40	1.50

Table(3.2,1,3,2)MLE & 95% I.E for the parameters of Model M_1

Parameters	MLE	95% Interval estimates
α	-2.425	-2.820 ; -2.030
β_1	-4.842	-5.417 ; -4.267
β_2	-2.793	-3.334 ; -2.252
β_3	0.371	0.294 ; 0.448
β_4	-0.637	-0.826 ; -0.448
β_5	-0.055	-0.076 ; -0.034

Table(3.3,P_1,3,2) Individual APA(M_1,i)

i=1,2,...7	0.20	0.43	0.43	0.29	0.12	0.13	0.14
i=8,9,..14	0.13	0.12	0.22	0.32	0.47	0.39	0.86

(The overall APA is equal 0.35).

Table(3.4,P_1,3,2) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.14	0.15	0.07	0.07	0.04	0.11	0.16
i=8,9,10	0.14	0.12	0.11				

(The overall APA is equal 0.12).

When Q_1 is modeled and Eq(3) is the true and Kosambi is the assumed map function.

Table(3.5,P_1,3,2)Examples showing the errors made by model M_1

$\mu^*=\frac{n_1}{1000}$	Logit(μ^*)	n^*_l= Logit(μ^*) -0.12	n^*_u= Logit(μ^*) +0.12	$\frac{Exp(n^*_l)}{1+Exp(n^*_l)}$	$\frac{Exp(n^*_u)}{1+Exp(n^*_u)}$
0.021	-3.842	-3.962	-3.722	0.019	0.024
0.117	-2.021	-2.141	-1.901	0.105	0.130
0.239	-1.158	-1.278	-1.038	0.218	0.262
0.363	-0.562	-0.682	-0.442	0.336	0.391
0.686	0.781	0.661	0.901	0.660	0.711

Table(3.3,P_W,3,2) Individual APA(M_1,i)

i=1,2,...7	0.48	0.47	0.30	0.56	0.65	0.60	0.33
i=8,9,..14	0.57	0.46	0.55	0.40	0.34	0.48	0.51

(The overall APA is equal 0.49).

Table(3.4,P_W,3,2) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.50	0.53	0.31	0.62	0.75	0.71	0.35
i=8,9,10	0.65	0.46	0.62				

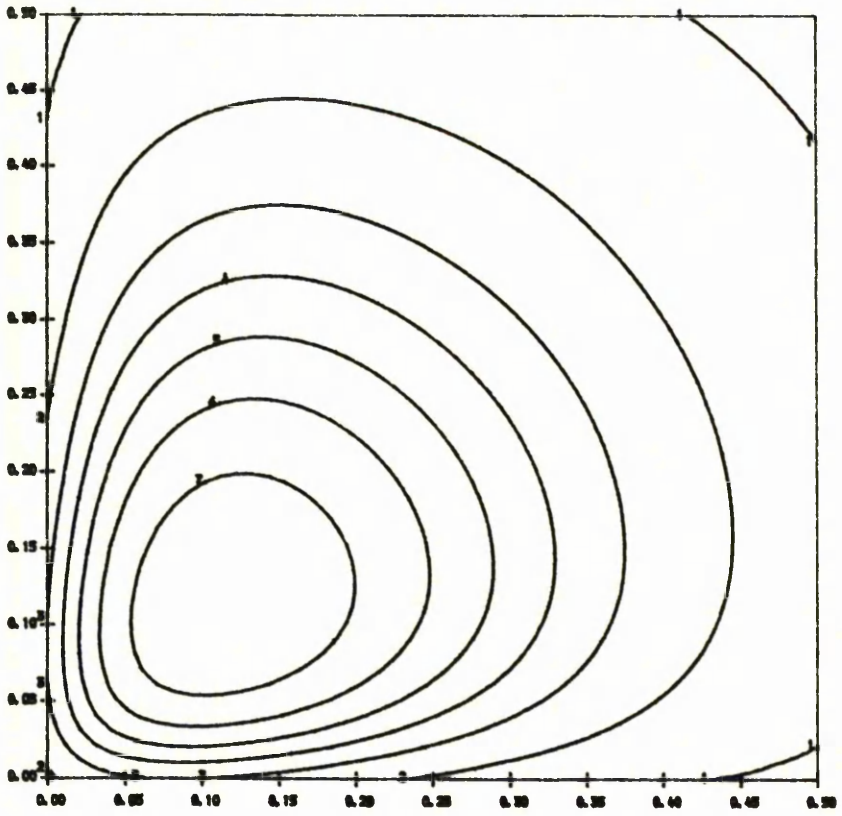
(The overall APA is equal 0.57).

Table(3.5,P_W,3,2)Examples showing the errors made by model M_1

$\mu^*=\frac{n_W}{1000}$	Logit(μ^*)	n^*_l= Logit(μ^*) -0.57	n^*_u= Logit(μ^*) +0.57	$\frac{Exp(n^*_l)}{1+Exp(n^*_l)}$	$\frac{Exp(n^*_u)}{1+Exp(n^*_u)}$
0.001	-6.907	-7.477	-6.337	0.001	0.002
0.004	-5.517	-6.087	-4.947	0.002	0.007
0.007	-4.955	-5.525	-4.385	0.004	0.012
0.014	-4.255	-4.825	-3.685	0.008	0.024

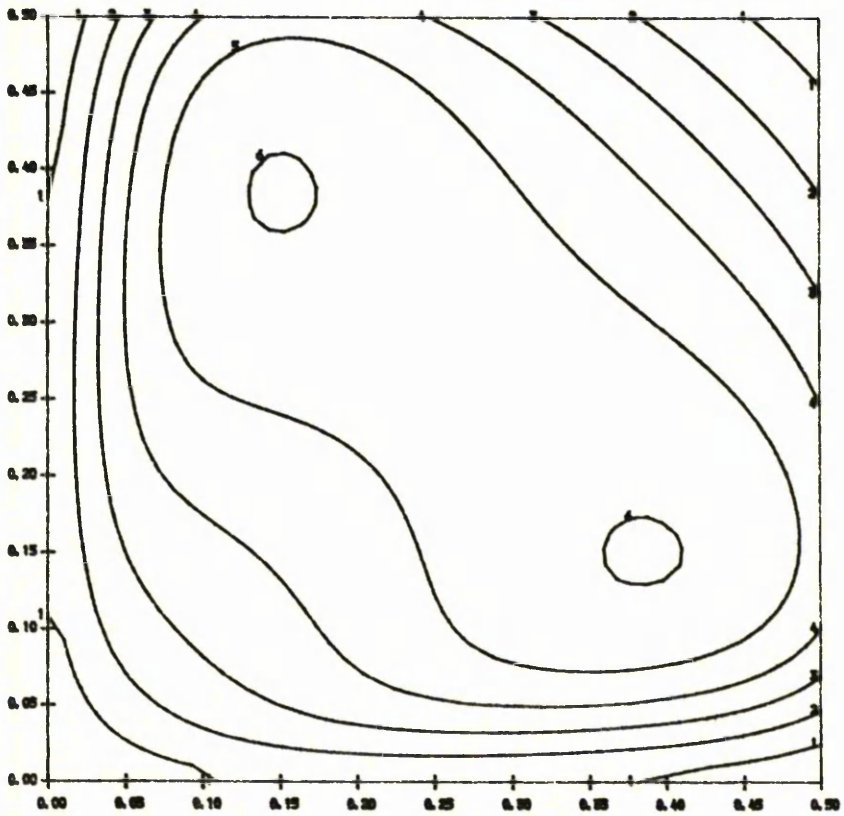
When Q_1 is modeled & Eq(3) is the true & Kosambi is the assumed map function

FIGURE(3.5,P1,3,2) Estimated contour plot of P_1 vs θ_1 & θ_2



HEIGHT AT
contour (1) = 0.001
contour (2) = 0.05
contour (3) = 0.15
contour (4) = 0.25
contour (5) = 0.35
contour (6) = 0.45
contour (7) = 0.55

FIGURE(3.5,PW,3,2) Estimated contour plot of P_w vs θ_1 & θ_2



HEIGHT AT
contour (1) = 0.002
contour (2) = 0.004
contour (3) = 0.006
contour (4) = 0.008
contour (5) = 0.010
countour (6) = 0.012

EQ(3) IS THE TRUE &

THE ASSUMED MAP FUNCTION

When Q_2 is modeled & Eq(3) is the true & the assumed map function

FIGURE (3.1,2,3,3) Discret contour plot of the data vs θ_1 & θ_2

```
0.50+  AAAAA   A   A   A   A   A   A   A   A   *
   -   AAAAA   A   A   A   A   A   A   A   A   A
   -   AAAAA   B   B   B   A   A   A   A   A   A
   -   AABBB   C   C   C   B   B   A   A   A   A
   -   ABBBC   D   D   D   C   B   B   A   A   A
0.25+  ABCDD   E   F   E   D   C   B   A   A   A
   -   BCCEE   F   F   F   E   D   C   B   A   A
   -   BCDEF   G   G   F   F   E   C   B   B   A
   -   BCDEE   G   G   F   E   D   C   B   A   A

0.050+ BCCDE   E   E   E   D   C   B   B   A   A
   -   ABCCD   E   E   D   C   B   B   A   A   A
   -   ABBCC   D   D   D   C   B   B   A   A   A
   -   ABBBC   C   C   C   B   B   A   A   A   A
   -   AAABB   B   B   B   A   A   A   A   A   A
0.000+
      +---------+---------+---------+---------+---------+-+
     0.00      0.10      0.20      0.30      0.40      0.50
```

* = Q2=1.000
A = 0.900---0.999
B = 0.800---0.899
C = 0.700---0.799
D = 0.600---0.699
E = 0.450---0.599
F = 0.300---0.449
G = 0.200---0.399

FIGURE (3.2,2,3,3) Logit of the data vs Z_{2j}

```
6.0+                                        *2
   -     3   *                              332
   -     2   2 22*                      *24358
   -     7   4 333     3 2 4 34555*
   -     2   7 766     3 4 4 5653
0.0+           234     5 5 4 52
   -                   3 3 2
      ---------+---------+---------+
             -4.0       -2.0       0.0
```

FIGURE (3.3,2,3,3) Logit of the data vs Z_{2j} for some Z_{1i}

When Q_2 is modeled and Eq(3) is the true and the assumed map function.

Table(3.1,2,3,3)A summary of the performance of different models

Model	Deviance	D.f	T.S	APA	$\frac{APA(M_1)}{APA(M_i)}$
M_0	41420.0	194		1.57	
M_1	492.1	189	15.6	0.20	1.00
M_2	242.1	135	6.5	0.14	1.43
M_t	73.7	91	-1.36	0.09	2.22

Table(3.2,2,3,2)MLE & 95% I.E for the parameters of Model M_1

Parameters	MLE	95% Interval estimates
α	6.820	6.727 ; 6.913
β_1	4.192	4.123 ; 4.261
β_2	1.204	1.181 ; 1.227
β_3	-0.018	-0.019 ; -0.017
β_4	-0.019	-0.024 ; -0.014
β_5	0.009	0.006 ; 0.012

Table(3.3,P_4,3,3) Individual APA(M_1,i)

i=1,2,...7	0.32	0.14	0.13	0.10	0.16	0.14	0.17
i=8,9,..14	0.14	0.17	0.03	0.22	0.20	0.16	0.42

(The overall APA is equal 0.20).

Table(3.4,P_4,3,1) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.20	0.16	0.17	0.11	0.09	0.13	0.19
i=8,9,10	0.14	0.12	0.12				

(The overall APA is equal 0.14).

Table(3.5,P_4,3,3)Examples showing the errors made by model M_1

$\mu^*=\frac{n_4}{1000}$	Logit(μ^*)	$n^*_l=$ Logit(μ^*) -0.14	$n^*_u=$ Logit(μ^*) +0.14	$\frac{Exp(n^*_l)}{1+Exp(n^*_l)}$	$\frac{Exp(n^*_u)}{1+Exp(n^*_u)}$
0.235	-1.180	-1.320	-1.040	0.211	0.261
0.560	0.241	0.101	0.381	0.525	0.594
0.722	0.954	0.814	1.094	0.693	0.749
0.842	1.673	1.533	1.813	0.822	0.860
0.970	3.476	3.336	3.616	0.966	0.974

When Q_2 is modeled & Eq(3) is the true & the assumed map function

FIGURE (3.4,2,3,3) Pearson residuals vs the fitted values

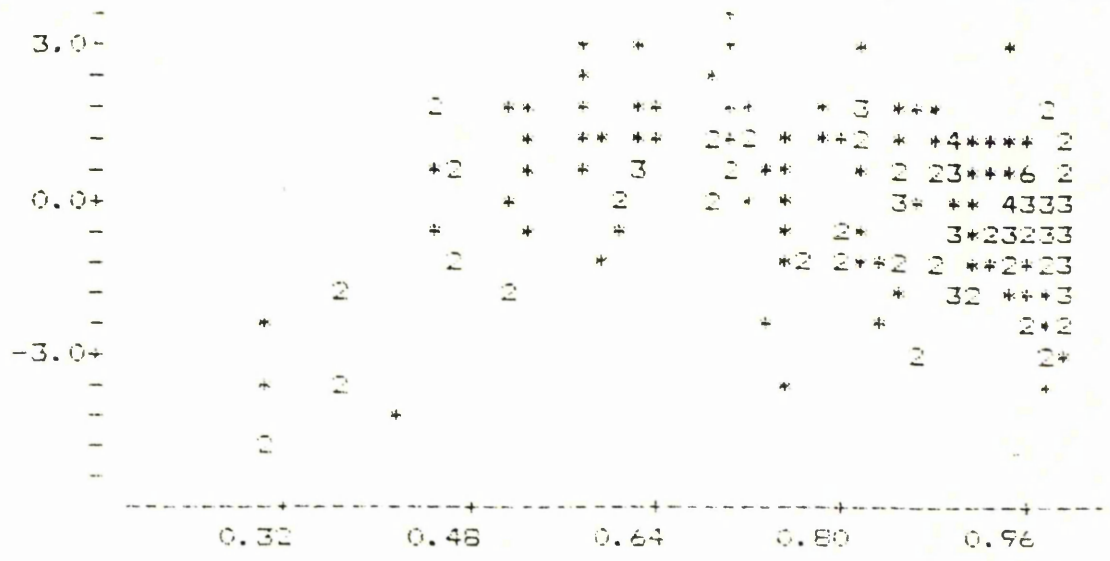
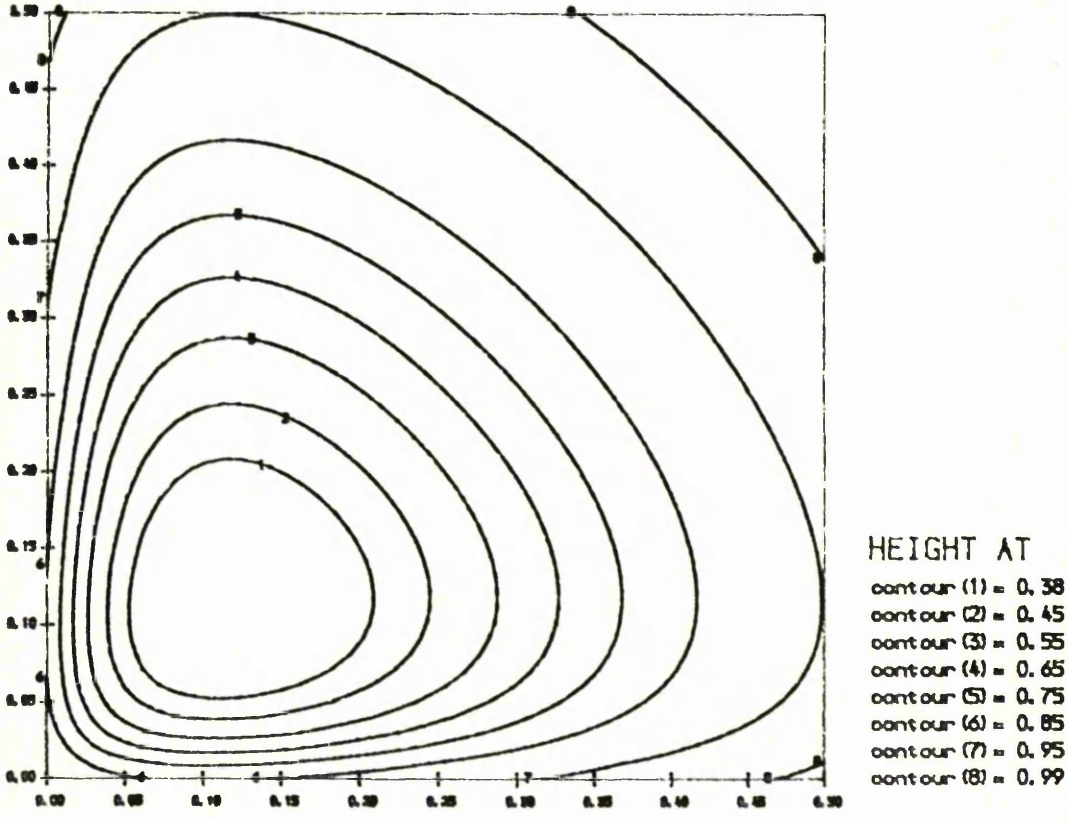


FIGURE (3.5,2,3,3) Estimated contour plot of P_4 vs θ_1 & θ_2



When Q_1 is modeled & Eq(3) is the true & the assumed map function

FIGURE (3.3,1,3,3) Logit of the data vs Z_{2j} for some Z_{1i}

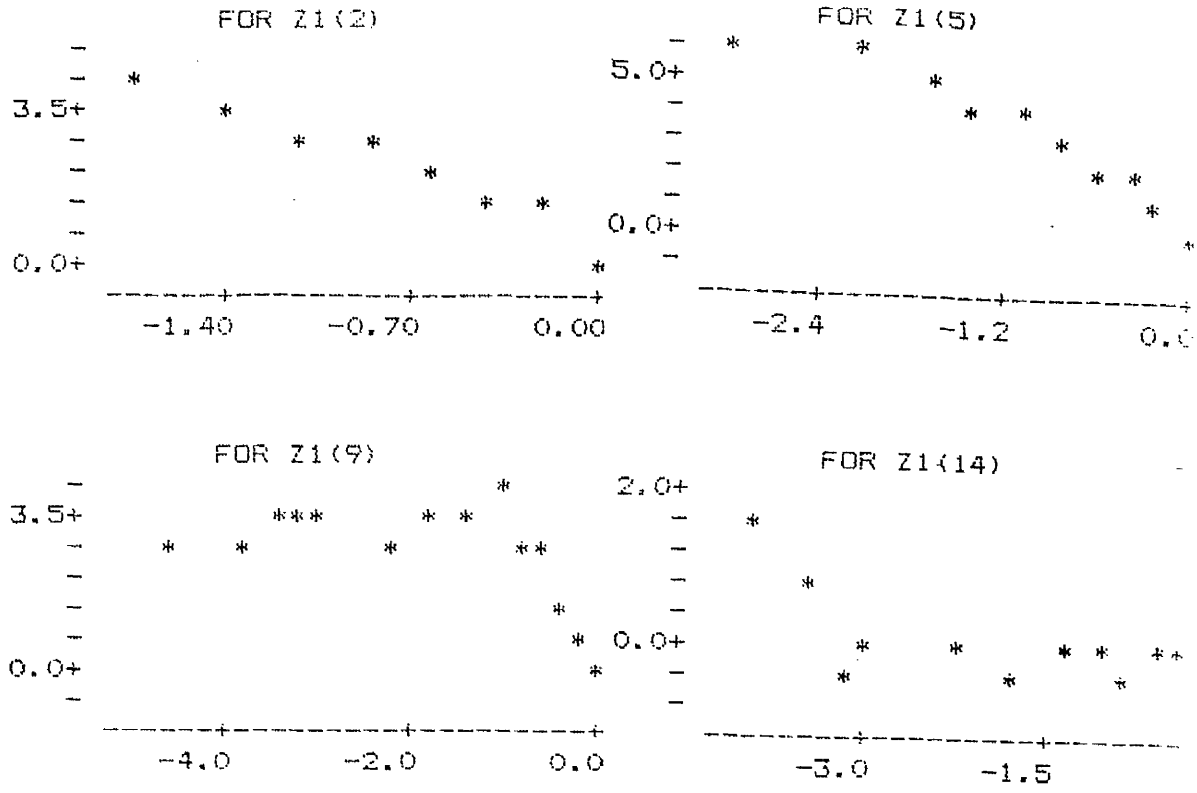
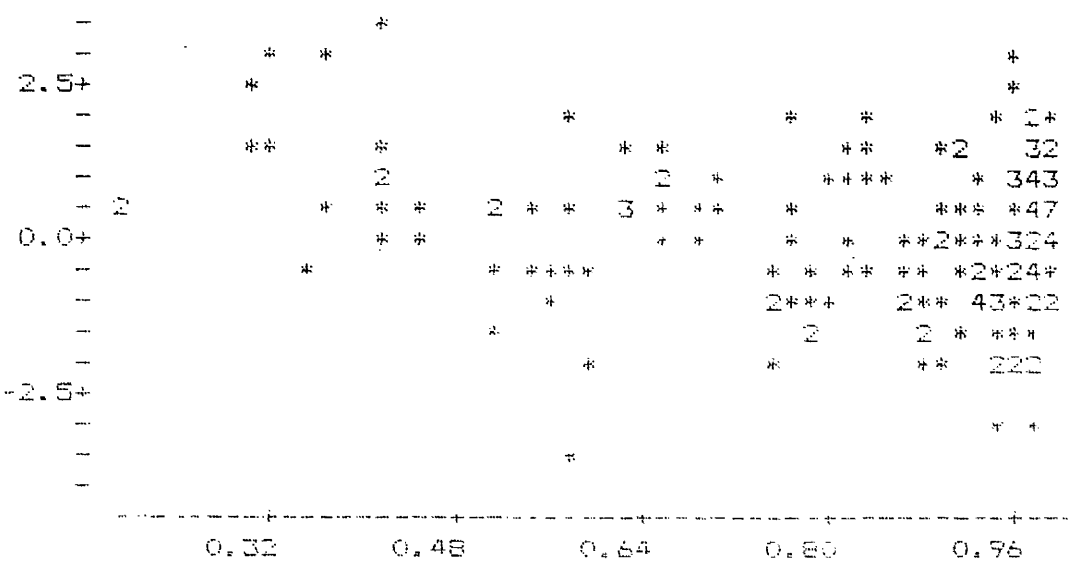


FIGURE (3.4,1,3,3) Pearson residuals vs the fitted values



When Q_1 is modeled & Eq (3) is the true & the assumed map function

FIGURE (3.1,1,3,3) Discret contour plot of the data vs θ_1 & θ_2

```
 0.50+  CDCCC      D     D     D     C     D     D     E     C     *
      -  DFDDE      E     E     E     D     E     D     C     B     H
      -  EFFFF      F     F     F     F     E     E     E     D     D
      -  FFFFG      G     G     G     G     G     F     G     E     D
      -  FGGGG      G     G     G     G     G     F     G     E     C
 0.25+  GGGGG      G     G     G     G     G     G     F     E     D
      -  GGGGG      G     G     G     G     G     G     F     E     C
      -  GGGGG      G     G     G     G     G     G     F     E     C
      -  GHGGG      G     G     G     G     G     G     F     D     D

0.050+  HHGHG      H     G     G     G     G     F     F     D     C
      -  HHGHH      G     G     G     G     F     G     F     E     C
      -  HHHGH      G     G     G     G     G     G     E     E     D
      -  HHHHH      G     G     G     G     G     F     E     D     F
      -  HHHMH      H     H     G     G     G     E     F     G     A
0.000+
       +-----------+-----------+-----------+-----------+-----------+--
      0.00        0.10        0.20        0.30        0.40        0.50
```

A = 0.0
B = 0.001---0.249
C = 0.250---0.499
D = 0.500---0.699
E = 0.700---0.799
F = 0.800---0.899
G = 0.900---0.999
H = 1.000
* = 00 = 1.000

FIGURE (3.2,1,3,3) Logit of the data vs Z_{2j}

```
 6.0+                 1+2     2
      -          2  ***     5 6 3  *
      -          *  232     * 4 6 8*
      -     4    2  233     2 2 2 39+3
      -     1    3  3**     2 * 2 2229+**
 0.0+            *   **     * * * ***+22+
      -                                *2
      -
       ------+-----------+-----------+
            -4.0        -2.0         0.0
```

When Q_1 is modeled and Eq(3) is the true and the assumed map function.

Table(3.1,1,3,3)A summary of the performance of different models

Model	Deviance	D.f	T.S	APA	$\frac{APA(M_1)}{APA(M_i)}$
M_0	3145.0	167		1.70	
M_1	248.6	162	4.8	0.55	1.00
M_2	143.0	133	0.6	0.41	1.35
M_t	90.0	76	1.1	0.37	1.49

Table(3.2,1,3,3)MLE & 95% I.E for the parameters of Model M_1

Parameters	MLE	95% Interval estimates
α	-2.209	-2.539 ; -1.879
β_1	-4.427	-4.871 ; -3.983
β_2	-2.579	-2.985 ; -2.173
β_3	0.368	0.306 ; 0.430
β_4	-0.590	-0.732 ; -0.448
β_5	-0.050	-0.066 ; -0.034

Table(3.3,P_1,3,3) Individual APA(M_1,i)

i=1,2,...7	0.18	0.30	0.14	0.08	0.20	0.14	0.16
i=8,9,..14	0.13	0.14	0.21	0.30	0.37	0.30	0.65

(The overall APA is equal 0.27).

Table(3.4,P_1,3,3) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leq 0.3$.

i=1,2,...7	0.13	0.16	0.07	0.12	0.07	0.13	0.21
i=8,9,10	0.14	0.12	0.14				

(The overall APA is equal 0.13).

When Q_1 is modeled and Eq(3) is the true and the assumed map function.

Table(3.5,P_1,3,3)Examples showing the errors made by model M_1

$\mu^*=\frac{n_1}{1000}$	Logit(μ^*)	n^*_l= Logit(μ^*) -0.13	n^*_u= Logit(μ^*) +0.13	$\frac{Exp(n^*_l)}{1+Exp(n^*_l)}$	$\frac{Exp(n^*_u)}{1+Exp(n^*_u)}$
0.030	-3.476	-3.606	-3.346	0.026	0.034
0.153	-1.711	-1.841	-1.581	0.137	0.171
0.278	-0.954	-1.084	-0.824	0.253	0.305
0.432	-0.274	-0.404	-0.144	0.400	0.464
0.755	1.125	0.995	1.255	0.730	0.778

Table(3.3,P_W,3,3) Individual APA(M_1,i)

i=1,2,...7	0.71	0.44	0.27	0.32	0.63	0.43	0.28
i=8,9,..14	0.54	0.41	0.38	0.28	0.51	0.59	0.72

(The overall APA is equal 0.48).

Table(3.4,P_W,3,3) Individual APA(M_1,i) within the restricted area of θ_1 & $\theta_2 \leqslant 0.3$.

i=1,2,...7	0.69	0.47	0.30	0.39	0.78	0.51	0.32
i=8,9,10	0.61	0.38	0.39				

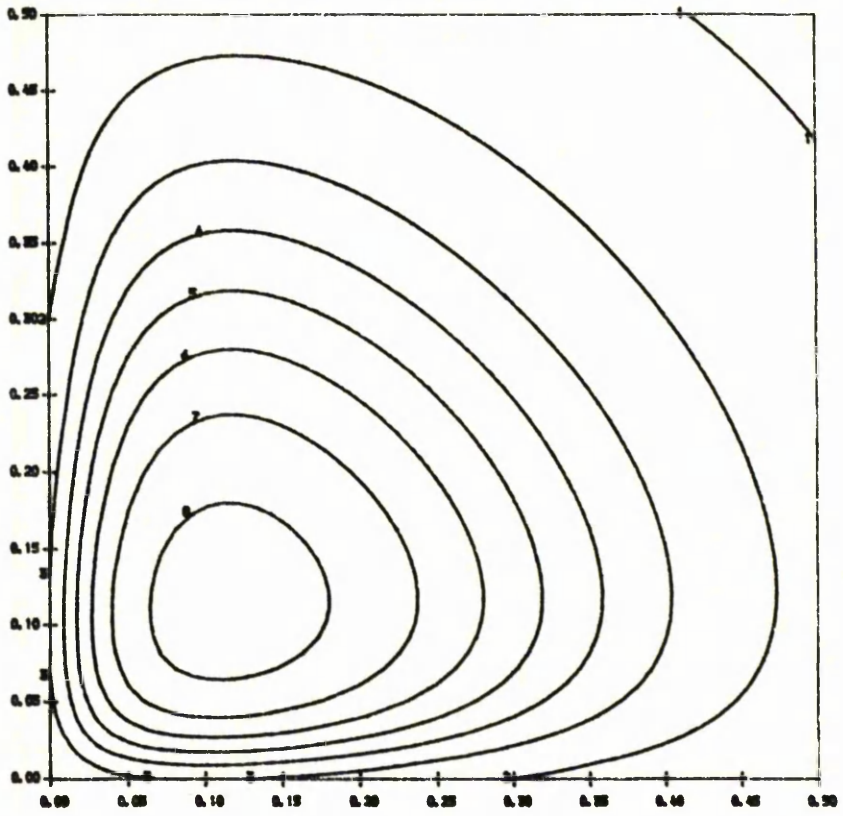
(The overall APA is equal 0.49).

Table(3.5,P_W,3,3)Examples showing the errors made by model M_1

$\mu^*=\frac{n_W}{1000}$	Logit(μ^*)	n^*_l= Logit(μ^*) -0.49	n^*_u= Logit(μ^*) +0.49	$\frac{Exp(n^*_l)}{1+Exp(n^*_l)}$	$\frac{Exp(n^*_u)}{1+Exp(n^*_u)}$
0.007	-4.955	-5.445	-4.465	0.004	0.011
0.011	-4.499	-4.989	-4.009	0.007	0.018
0.025	-3.664	-4.154	-3.174	0.015	0.040

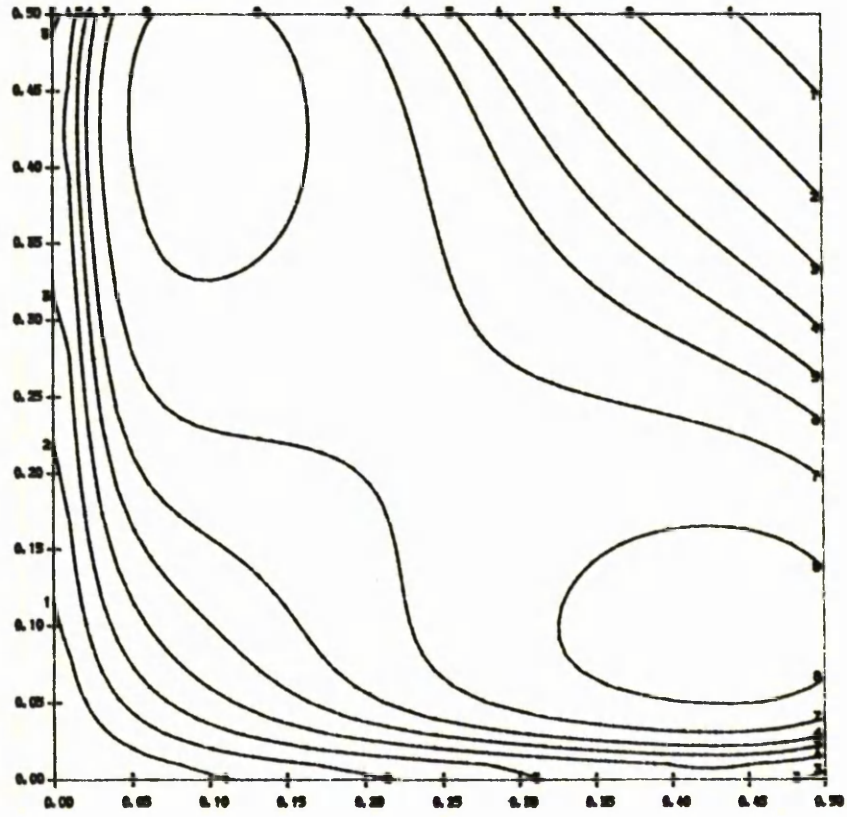
When Q_1 is modeled & Eq(3) is the true & the assumed map function

FIGURE(3.5,P1,3,3) Estimated contour plot of P_1 vs θ_1 & θ_2



HEIGHT AT
contour (1) = 0.001
contour (2) = 0.05
contour (3) = 0.15
contour (4) = 0.25
contour (5) = 0.35
contour (6) = 0.45
contour (7) = 0.55
contour (8) = 0.65

FIGURE(3.5,PW,3,3) Estimated contour plot of P_w vs θ_1 & θ_2



HEIGHT AT
contour (1) = 0.002
contour (2) = 0.004
contour (3) = 0.006
contour (4) = 0.008
contour (5) = 0.010
contour (6) = 0.012
contour (7) = 0.015
contour (8) = 0.020

CHAPTER FOUR: Comparing approximate intervals using the likelihood approach

(4.1) Introduction

The aim of the following two chapters is to carry out a comparative study between different ways of constructing interval estimates in genetics. The next chapter, chapter 5, will deal with a simple comparison between the likelihood and Bayesian approaches in constructing interval estimate , for short IE, for the most simplistic model in genetics dealing with a two loci situation and a phase known or unknown double backcross mating.

This chapter, in contrast, will be concerned with the more complicated case of a three loci situation and a phase known or unknown triple backcross mating. Interval estimates will be constructed and compared using only the likelihood approach but in conjunction with certain methods of approximation for the likelihood function. Actually constructing IE will demand either the knowledge of the original likelihood or an approximation of it. Two methods of approximation are compared. The first was introduced by Ott and therefore we shall call it Ott approximation, whereas the second is introduced in this study and uses Least square method to find the approximation and therefore is called the LSM approximation.

(4.2) Notation, mating type and map function

As seen in chapter 1, section(1.6.1)-C, under three loci situation with two codominant alleles each denoted by (A,a),(B,b) and (C,c) respectively arranged in that order and a phase known triple backcross mating of the form ABC/abc×abc/abc, the number of the different types of offsprings produced by this mating are random variables having the multinomial distribution with

parameter N, as the total number of offsprings and corresponding probabilities of success for each category or type of offspring. Actually under this situation and by using the same notation of chapter one, a data point $\underline{r}=(r_1r_2r_3r_4)$ will have the multinomial distribution described in (1.14). Whereas if we had a number of phase unknown matings with two offsprings each and such that the total number of offsprings is N the multinomial distribution (1.15) should be used instead. Under both situations α, β, γ and δ are functions of θ_{ab}, θ_{bc} and θ_{ac} as stated in (1.6) and where these later parameters will be subject to the restriction(1.11). In this chapter let for simplicity θ_{ab}, θ_{bc} and θ_{ac} be denoted by θ_1, θ_2 and θ_{1+2} respectively. We have also seen before that by using a certain map function the dimensionality of the problem will be reduced from three to two unknown parameters θ_1 and θ_2, whereas θ_{1+2} could be written as function of these two parameters using formula (1.21) and therefore (appendix(2.1)) restriction(1.11) will be reduced to just

$0 < \theta_1 < 0.5$

$0 < \theta_2 < 0.5$

Morton Eq(3) map function, formula(1.25), was one of the most recent map functions mentioned in chapter one, which when compared to previous ones in fitting real multipoint data, Eq(3) gave the best fit (see page 39). According to this result Eq(3) has been chosen for this part of the study.

(4.3) Constructing IE

(4.3.1) Introduction

The likelihood approach has been adopted for constructing IE for the parameters of interest $\underline{\theta}=(\theta_1\ \theta_2)$. If $l(\underline{\theta})$ is the ln likelihood function of $\underline{\theta}$, then a joint IE based on the likelihood approach will have the following general form:

$$IE(\underline{\theta},h) = \{\underline{\theta} \text{ such that } L(\underline{\theta})>hL(\hat{\underline{\theta}})\}$$

$$= \{\underline{\theta} \text{ such that } l(\underline{\theta})>h'+l(\hat{\underline{\theta}})\} \qquad (4.1)$$

where $h'= \ln(h)$ and $\hat{\underline{\theta}}$ is the MLE of $\underline{\theta}$. Also, note that the IE depends on the value of the constant h to determine the required confidence coefficient c, (i.e) h is chosen such that $\text{Prob}\{\underline{\theta}_t \in I(h)\} = c$, where $\underline{\theta}_t$ is the true parameter vector.

The requirement of an exact confidence coefficient, for short CC, can only be satisfied in special cases. Therefore, the following large sample result will be useful in constructing aprroximate CC. Under regularity conditions

$$l(\underline{\theta}_t) - l(\hat{\underline{\theta}}) \sim -(0.5)x^2(p) \qquad (4.2)$$

where p is the number of unknown parameters, (i.e) here p=2. Now, choosing $h'=-(0.5)x^2(p;c)$ will provide us with a recipe for obtaining an IE for $\underline{\theta}$ based on the likelihood and with approximate CC equal c.

To construct an IE for a subset of the parameter $\underline{\theta}$, say θ_1, we will use the idea of the profile ln likelihood, $l(\theta_1\ \theta_2^*(\theta_1))$, defined by

$$l(\theta_1\ \theta_2^*(\theta_1)) = \max_{\theta_2} l(\underline{\theta})$$

Therefore an IE for θ_1 would be:

$$I(\theta_1,h) = \{\theta_1 \text{ such that } l(\theta_1\ \theta_2^*(\theta_1))-l(\hat{\underline{\theta}})>h\} \qquad (4.3)$$

Again, by using large sample results and under regularity condition, h could be chosen as equal to $-(0.5)x^2(1,c)$ to ensure an approximate CC equal c.

For large n, it will often be true that $l(\underline{\theta})$ can be well approximated in the neighbourhood of $\hat{\underline{\theta}}$ by a quadratic function in $\underline{\theta}$. This can be done by expanding $l(\underline{\theta})$ to two terms in a Taylor expansion about $\hat{\underline{\theta}}$, (i.e)

$$l(\underline{\theta}) = l(\hat{\underline{\theta}}) + (\underline{\theta}-\hat{\underline{\theta}})\left.\frac{\partial l}{\partial \underline{\theta}}\right|_{\underline{\theta}=\hat{\underline{\theta}}} + \frac{1}{2}(\underline{\theta}-\hat{\underline{\theta}})^T\{\left.\frac{\partial^2 l}{\partial \underline{\theta}^2}\right|_{\underline{\theta}=\hat{\underline{\theta}}}\}(\underline{\theta}-\hat{\underline{\theta}}) \qquad (4.4)$$

(Note that the second term in (4.4) will vanish if $\hat{\underline{\theta}}$ does not occur at the boundaries). This is equivalent to saying that for large n the likelihood of $\underline{\theta}$ could be approximated by a normal likelihood. In this part of the study we are going to use three methods of constructing IE for $\underline{\theta}$, θ_1 or θ_2. The difference between them depends on the actual ln likelihood used in (4.1) and (4.3). The first method is:

(1)Original likelihod: As the name suggests, the original likelihood of $\underline{\theta}$ will be used to find these intervals.

The following two methods are based on the above large sample approximation:

(2)Ott quadratic approximation: This method is mentioned by Ott (1985 page88-95). He argued that in practice the likelihood may be taken to approximately represent a bivariate normal distribution, so that the likelihood will be quadratic in θ_1 and θ_2. The contour lines of $l(\underline{\theta})$ would be represented by a number of ellipses and therefore $l(\underline{\theta})$ could be written as follows:

$$l(\underline{\theta}) = a_1\theta_1^2 + a_2\theta_2^2 + b_1\theta_1 + b_2\theta_2 + c\theta_1\theta_2 + d$$

Any six points $\underline{\theta}_i$ and their corresponding ln likelihood, denoted by $l_i=l(\underline{\theta}_i)$ for i=1,2,..6, can determine the unknown coefficients a_1, a_2, b_1, b_2, c and d in the above function. Ott suggested the lay out produced in table(4.1) for the choice of these six points and emphasized that the results will be most accurate if l_3 corresponds to the largest ln likelihood, although it is not a requirement.

Table(4.1)

	θ_{11}	θ_{12}	θ_{13}
θ_{21}		l_1	
θ_{22}	l_2	l_3	l_4
θ_{23}		l_5	l_6

He also stated that if we believe that the ln likelihood is skewed then it is important to use a transformation of the parameters. Strict techniques could be used to find the best transformation suggested by the data. For example among all power transformation $x= \theta^{\lambda}$ and under a certain mating type, choose λ which will lead to a zero third and higher derivative of the ln likelihood at $\hat{x}$. But in practice, he suggested that it would be simpler generally to apply a specific, rather mild transformation such as $\phi=\sqrt{\theta}$. For convenience, this transformation will be adopted by us in this study. The recipe of finding a joint interval $\underline{\theta}$ and a marginal IE for θ_i under this method will be:

(i)Using the six chosen points, find the coefficient of:

$$Q1O(\underline{\phi}) = a_1\phi_1^2 + a_2\phi_2^2 + b_1\phi_1 + b_2\phi_2 + c\phi_1\phi_2 + d \qquad (4.5)$$

where $Q1O(\underline{\phi})$ is the quadratic ln likelihood given the Ott approximation.

(ii)A 95% IE for $\underline{\phi}$ is:

$$IE(\underline{\phi},h) = \{\underline{\phi} \text{ such that } Q1O(\underline{\phi})-Q1O(\hat{\underline{\phi}}_O)>h\} \qquad (4.6,a)$$

where $h= -(0.5)\chi^2(2,0.95)$ and $\hat{\underline{\phi}}_O$ is the MLE of $\underline{\phi}$ using $Q1O(\underline{\phi})$.

(iii)A 95% IE for ϕ_i where i=1,2 is:

$$IE(\phi_i,h') = \{\phi_i \text{ such that } Q1O(\phi_i\ \phi_j^*(\phi_i)) - Q1O(\hat{\underline{\phi}}_O)>h'\} \qquad (4.7,a)$$

for j=1 or 2 and $i\neq j$ and when $h'= -(0.5)\chi^2(1,0.95)$.

(iv)Since $\underline{\phi}$ is a 1-1 transformation of $\underline{\theta}$ when $0<\theta<0.5$, then the intervals defined by:

$$IE(\underline{\theta},h) = \{\underline{\theta} \text{ such that } \underline{\phi}\in I(\underline{\phi},h)\} \qquad (4.6,b)$$

$$IE(\theta_i,h') = \{\theta_i \text{ such that } \phi_i\in I(\phi_i,h')\} \qquad (4.7,b)$$

will have an approximate 95% CC.

(3)<u>LSM quadratic approximation</u>: Instead of using six points near the maximum to determine the unknown coefficients in (4.5) use many points as near as possible to the 95% contour of the original likelihood and then fit the quadratic likelihood to

these points using Least square method, hence the name of this method; the resultant quadratic likelihood will be denoted by $QlL(\underline{\Phi})$. The recipe of finding the required IE will be very similar to that when using the Ott approximation apart from (i) for obvious reason. Also as we suspect that this method will not give a good approximation to the MLE, then when using (4.6,a) and (4.7,a) use the evaluation of the LSM quadratic likelihood at $\hat{\underline{\Phi}}_o$, the MLE of the Ott approximation. This means that an approximate 95% IE for $\underline{\Phi}$ would be redefined as follows:

$$IE(\underline{\Phi},h) = \{\underline{\Phi} \text{ such that } QlL(\underline{\Phi})-QlL(\hat{\underline{\Phi}}_o)>h\} \qquad (4.6,c)$$

And an approximate 95% marginal IE for Φ_i would be redefined as follows:

$$IE(\Phi_i,h') = \{\Phi_i \text{ such that } QlL(\Phi_i\ \Phi_j^*(\Phi_i)) - QlL(\hat{\underline{\Phi}}_o)>h'\} \qquad (4.7,c)$$

(Note that the choice of the points under Ott and the LSM approximations will be discussed in the next section and later on in details).

(4.3.2) A prior comparison between Ott and LSM:

The advantage of using Ott approximation is actually to force our quadratic to have its maximum near the orignal one, although the precision of this method away from the maximum is not known. Therefore, we suspect that it could lead, in some cases, to a 95% contour away from the original one.

Using points near the 95% contour of the original likelihood in deriving the LSM intervals will probably lead to a nearer 95% IE to the original one, altough it could lead to a misleading MLE. Figure(4.1) below is a crude picture of our perception of a 95% contour for $\underline{\Phi}$ produced by Ott and LSM in comparison to using the original likelihood.

In order to use the LSM we have to choose points near the original 95% contour which, if known, will cancel out the urge

for an approximation. Instead, we will have to use points near to an approximate 95% contour. The LSM can only be used as a secondary approximation to a primary one, here our primary approximation will be the Ott approximation.

Figure (4.1)

95% contour using Ott, MLE is O

95% contour using LSM, MLE is L

95% original contour, MLE is *

(4.4) Assessing the performance of the methods

The assessment of the three methods will be based on calculating, for each method in turn, the exact CC and expected length, for short EL. These measurements would be calculated as follows. For a data vector $\underline{r}$ arising from either the multinomial distribution of (1.14) or (1.15), then the finite sample space R of $\underline{r}$ could be written as:

$$R = \{(0,0,0,N),(0,0,1,N-1),\ldots,(N,0,0,0)\}$$

For convenience let IE(l,m,$\underline{r}$) be a 95% approximate IE produced by method l for the parameter m and by using the data vector $\underline{r}$, where

l = 1, if the Original likelihood method is used.

= 2, if the Ott approximation is used.

= 3, if the LSM approximation is used.

m = 1, if a marginal IE is produced for θ_1.

= 2, if a marginal IE is produced for θ_2.

= 3, if a joint IE is produced for $\underline{\theta}$.

Now let

$$T(l,m,\underline{r}) = \begin{cases} 1 & \text{if } \underline{\theta}_t(m) \in IE(l,m,\underline{r}) \\ 0 & \text{otherwise} \end{cases}$$

Furthermore, let $LE(l,m,\underline{r})$ be the length of $IE(l,m,\underline{r})$ when $m=1,2$ only. Then, the exact CC and EL for method 1 and parameter m denoted by $CC(l,m)$ and $EL(l,m)$ respectively will be:

$$CC(l,m) = \sum_{\underline{r}\in R} T(l,m,\underline{r})p(\underline{r}|N,\underline{\theta}_t) \qquad (4.8)$$

$$EL(l,m) = \sum_{\underline{r}\in R} LE(l,m,\underline{r})p(\underline{r}|N,\underline{\theta}_t) \qquad (4.9)$$

where $p(\underline{r}|N,\underline{\theta}_t)$ is the probability distribution of $\underline{r}$ using either (1.14) or (1.15).

Our criteria for assessing the three methods would be to give more credit to the method with CC greater or equal 0.95. Among those methods satisfying the first criterion, the one with a shorter EL would be preferred when marginal intervals are compared. Furthermore, if we wanted to compare the two methods of approximations Ott and LSM in more detail, we can calculate the exact distribution of the difference between the intervals' lengths denoted by D(m) with m=1,2 only. Let:

$$D(m) = LE(2,m,\underline{r}) - LE(3,m,\underline{r})$$

D(.) is a random variable with sample space S(D), which is a finite subset of :

$$S = \{ d \text{ such that } -0.5<d<0.5\}$$

A histogram of the distribution of D(.) could be presented by categorising the sample space S. Table(4.2) shows our choice of the categories.

Knowledge of the true parameters $\underline{\theta}_t$ and N is crucial in order to calculate CC(.,.), EL(.,.) and the distribution of D(.). Restricting ourself to the choice of small values of $\underline{\theta}_t$ and

moderate number of N will give us an idea about the performance

Table(4.2)

Category C_i	$P(C_i)$
$-0.5<D(.)\leq-0.49$	$P(C_1)$
$-0.49<D(.)\leq-0.48$	$P(C_2)$
.	
.	
$0.49<D(.)\leq0.5$	$P(C_{100})$

$$\left[\text{where} \quad P(C_i) = \sum_{\substack{\text{all } r\in R \text{ such} \\ \text{that } D(.)\in C_i}} p(\underline{r}|N,\underline{\theta}_t) \qquad (4.10)\right]$$

of the different methods for the common and interesting situation for the geneticists. Therefore the choice of $\underline{\theta}_t$ and N has been, N=25 and θ_1 and θ_2 varying as in table(4.3)

Table(4.3)

θ_2 \ θ_1	0.05	0.1	0.15
0.05	×	×	×
0.1		×	×
0.15			×

(Note that only the combinations marked "×" have been considered because of the symmetry of the problem about θ_1 and θ_2).

(4.5) Practical consideration

(4.5.1)Simulation

When N=25, the total number of the data vectors $\underline{r}$ in R is equal to "3276". Considering all of these points will be time consuming especially if we understand that some points will be associated with very small probability . To overcome this problem, we decided to simulate data from the required multinomial distribution under the different combination of $\underline{\theta}_t$. Let R_s be the set of all data vectors $\underline{r}$ occuring in this simulation, then:

$$R = R_s \cup \bar{R}_s$$

If the number of simulated data, called I, is large enough then:

$$\sum_{\underline{r}\in R_S} p(\underline{r}|N,\underline{\theta}_t) \simeq 0$$

(The reader is referred to the last column of table(4.5), page 162, which provides numerical examples of the above statment. From which we can see that, when "I" was equal 10000 and for the defined $\underline{\theta}_t$, the above probability was at most equal 0.005).

Now, using all points occurring in R_S recalculate (4.8), (4.9) and (4.10) as:

$$CC(l,m) = \sum_{\underline{r}\in R_S} T(l,m,\underline{r})p(\underline{r}|N,\underline{\theta}_t) \qquad (4.8,a)$$

$$EL(l,m) = \sum_{\underline{r}\in R_S} LE(l,m,\underline{r})p(\underline{r}|N,\underline{\theta}_t) \qquad (4.9,a)$$

$$P(C_i) = \sum_{\substack{\text{all } \underline{r}\in R_S \text{ such} \\ \text{that } D(.)\in C_i}} p(\underline{r}|N,\underline{\theta}_t) \qquad (4.10,a)$$

(4.5.2)Approximation of the inverse map function

To calculate either the original ln likelihood of $\underline{\theta}$, $l(\underline{\theta})$, or the probability distribution of $\underline{r}$, $p(\underline{r}|N,\underline{\theta}_t)$, we have to determine the exact value of θ_{1+2} as a function of θ_1 and θ_2, where by using (1.21) $\theta_{1+2} = f^{-1}(f(\theta_1)+f(\theta_2))$. Actually from (1.25) $f(\theta)$ is equal to:

$$x= \frac{-1}{12}\ln\frac{(1-2\theta)^2}{(1+2\theta+4\theta^2)}+\frac{\sqrt{3}}{6}\tan^{-1}\frac{(1+4\theta)}{\sqrt{3}}-0.15115$$

By looking carefully at this function, it is easy to see that it does not have a direct mathematical inverse, although a numerical one could be calculated at any point $0<\theta<0.5$. Let $x=f(\theta)$, for $0<\theta<0.5$, then $\theta=f^{-1}(x)$ for $0<x<\infty$. By using a large number of points (θ_k,x_k) and fitting a cubic spline function to θ on x

(appendix(4.1)), we succeeded in producing a very good approximation to $f^{-1}(x)$. Figure(4.2) shows $f(\theta)$ and its approximated inverse $f^{-1}(x)$, which is a straight line curve passing throught the origin and with slope 1, also shown in the figure, red curve, the inverse of the approximated inverse $f(f^{-1}(x))$, it is very difficult to see this latter curve because of it coinciding exactly with the original $f(\theta)$.

(4.6) Application, given the phase known situation

(4.6.1) Introduction

This section deals with the application of the three methods of IE discussed in section (4.3) when using a data vector $\underline{r}$ arising from the multinomial distribution of (1.14).

(4.6.2) Original likelihood

(A) Joint interval:

If $\underline{r}$ comes from the multinomial distribution of (1.14), then by using formula (4.1) an approximate 95% joint IE for $\underline{\theta}$ would be defined if $h'=-(0.5)x^2(2,0.95)$ and where:

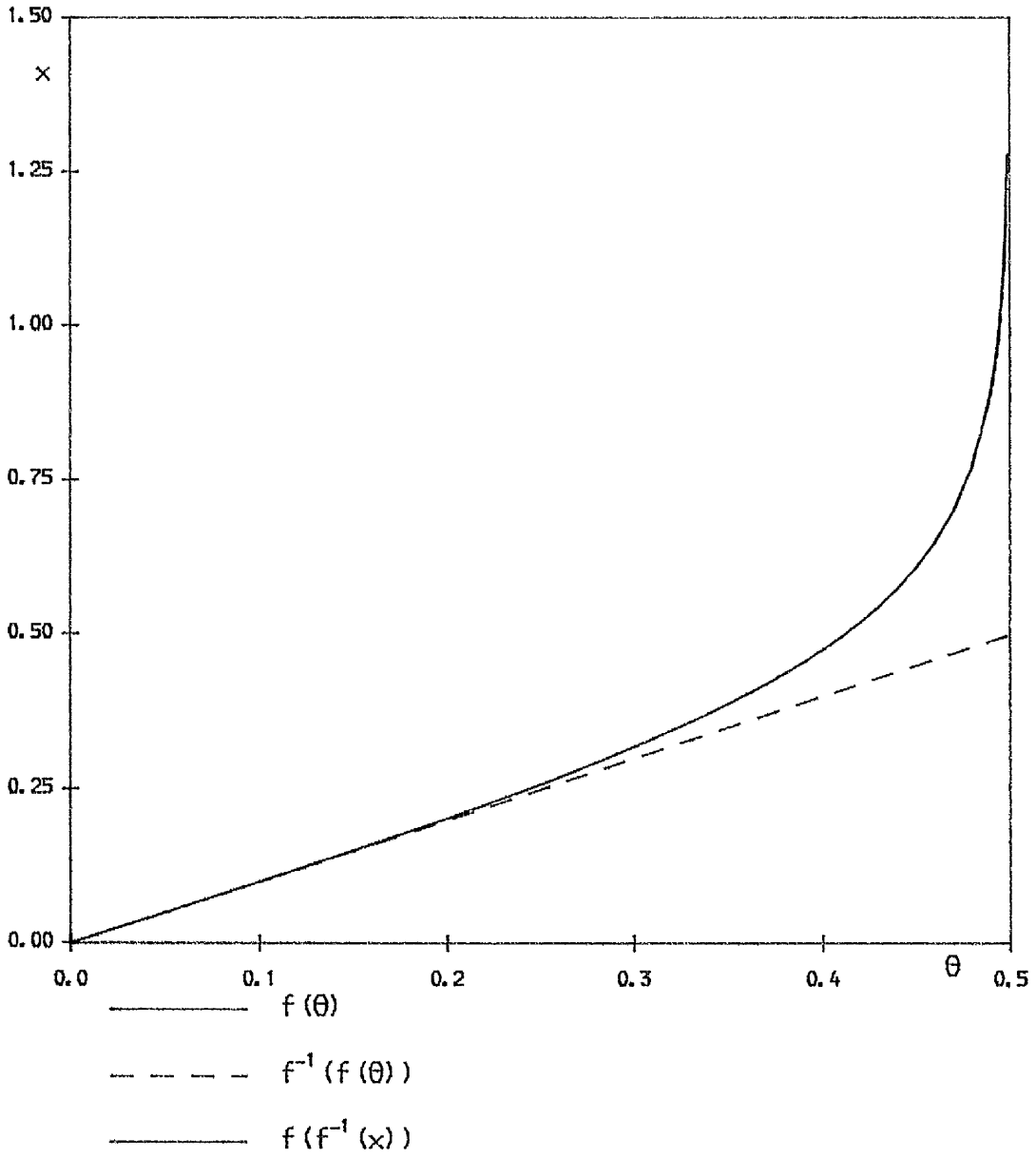
$$l(\underline{\theta}) = \text{constant} + r_1\ln(\theta_1+\theta_2-\theta_{1+2}) + r_2\ln(\theta_1-\theta_2+\theta_{1+2}) + r_3\ln(\theta_{1+2}-\theta_1+\theta_2) + r_4\ln(2-\theta_1-\theta_2-\theta_{1+2}) \qquad (4.11)$$

$$\text{for } 0 < \theta_1 < 0.5, \quad 0 < \theta_2 < 0.5 \quad \text{and} \quad \theta_{1+2}=f^{-1}(f(\theta_1)+f(\theta_2))$$

(For convenience we will call the area where $0<\theta_1<0.5$ and $0<\theta_2<0.5$ the feasible region).

Maximising the ln likelihood, $l(\underline{\theta})$ defined in (4.11) can not be done analytically, numerical methods have to be used. But if $l(\underline{\theta})$ is multimodal within the feasible region then using numerical methods could be misleading. Again, it is difficult to guarantee a unique maximum for $l(\underline{\theta})$ analytically. The only guarantee that we can provide is a plot of $L(\underline{\theta})$ under various choices of $\underline{r}$,

FIGURE (4.2) A plot of EQ(3), $x=f(\theta)$, & its approximated inverse vs θ



shown in figure(4.3)(a,b,c,d). From the plots, and others not shown here, we can see that $L(\underline{\theta})$ is probably both a concave and a unimodal function of $\underline{\theta}$, within the feasible region. Therefore using numerical maximisation will not be dangerous. A quasi-Newton method has been used with the aid of the NAG computer package.

Also note that some of the points $\underline{r}$ will have its MLE occuring at one or more of the boundaries of the feasible region, which means that for these points the regularity conditions that support the statment (4.2) will not hold. Nevertheless we decided to use the same recipe of the IE for these points as before, in order to see how the method, in general, is behaving. (The same comment will apply for both quadratic approximations).

(B)Marginal interval:

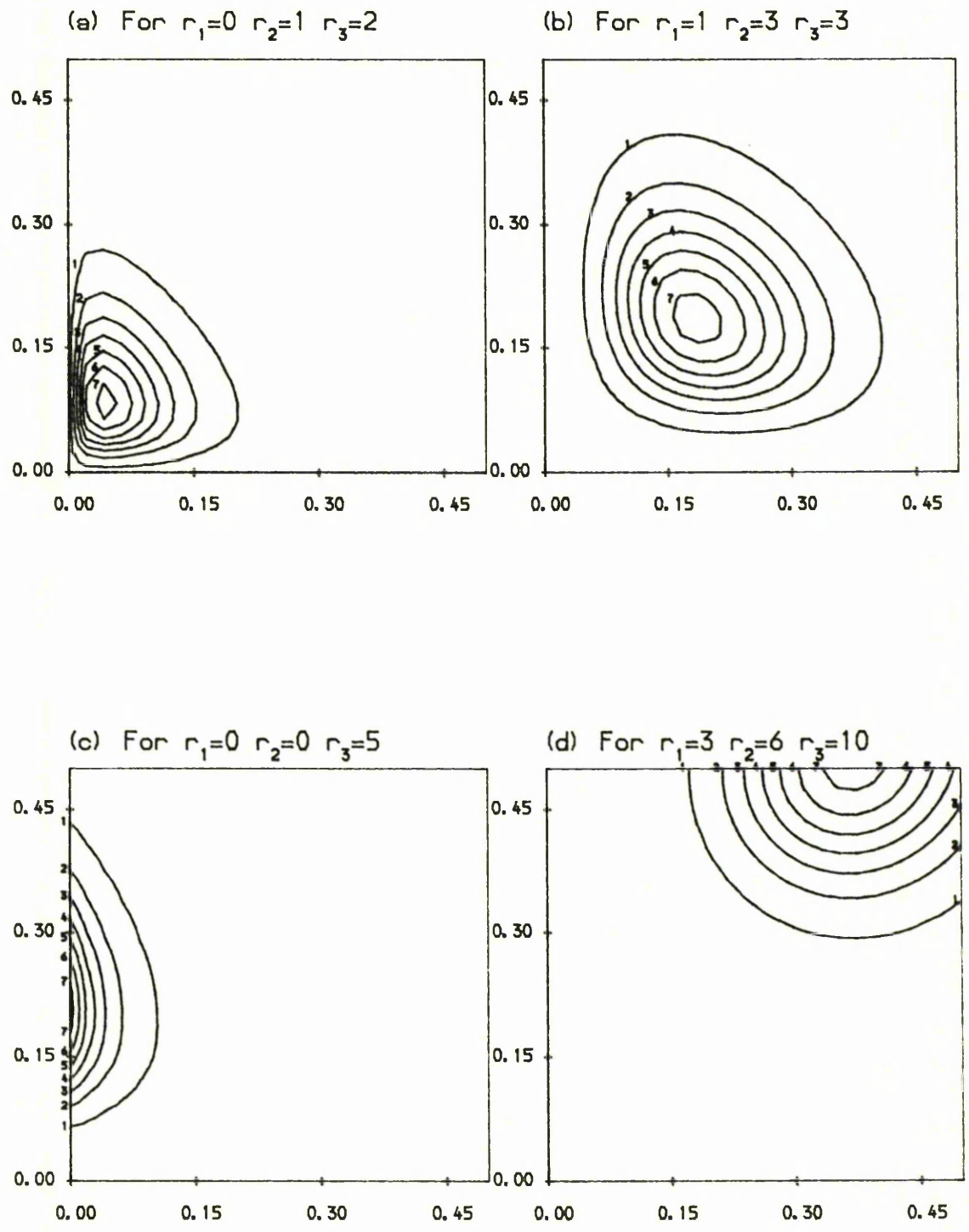
From formula (4.3) an approximate 95% IE for θ_1 would be defined if $h=-(0.5)x^2(1,0.95)$. Therefore, the set of points of θ_1 that satisfy (4.3), are all the values of θ_1 between the lower and upper roots of the following function:

$$G(\theta_1) = l(\theta_1\ \theta_2^*(\theta_1)) - l(\hat{\underline{\theta}}) + 1.92$$

provided that the roots are within the region of [0.0,0.5]. Numerical methods have to be adopted again to find $LL(\theta_1)$ and $UL(\theta_1)$, the lower and upper limit of IE(1,1) respectively. The profile ln likelihood, $l(\theta_1\ \theta_2^*(\theta_1))$, can not be written as an explicit function of θ_1 because θ_2^* can only be found numerically. But, if $l(\underline{\theta})$ is both concave and unimodal, then $l(\theta_1\ \theta_2^*(\theta_1))$ will be unimodal too. A full proof of this last statment is not supplied here but an intuitive and logical argument is. Figure(4.4)(a) is a contour plot of the joint ln likelihood function; at a fixed point θ_1^0, $\theta_2^*(\theta_1^0)$ could be found from the plot by drawing a vertical line parallel to the y axis at the

PHASE KNOWN SITUATION

FIGURE (4.3) Contour plot of the likelihood when N=25



value θ_1^0. But because of the concavity of the contours, this line will only be tangent to one special contour, the tangent point will be $\theta_2^*(\theta_1^0)$, points A, B and C on the plot for example. Figure(4.4)(b) is an incomplete plot of the profile ln likelihood, where A', B' and C' are the corresponding points of A, B and C of the joint ln likelihood. Our aim is to show that neither before nor after the maximum of the profile ln likelihood which occurs at A', the imaginary situation which corresponds to the lay out of the points B', D' and C'can not occur. This could easily be seen from figure (4.4)(a) where between the two lines bb' and cc', and because of the concavity of the likelihood, all the contours, which occur between the straight line joining B and C, will have a height higher than the height at C and lower than the height at B. Thus by using the unimodality of the profile likelihood, numerical methods could be used to determine both $LL(\theta_1)$ and $UL(\theta_2)$. Figure(4.5) is a flow chart showing the essential steps involved in determining $LL(\theta_1)$. A brief description of the method is as follows:

(i)If $\underline{\hat{\theta}}=(\hat{\theta}_1\ \hat{\theta}_2)$ is the MLE of $\underline{\theta}$, then the maximum of $G(\theta_1)$ occurs at $\hat{\theta}_1$; therefore $0 \leqslant LL(\theta_1) \leqslant \hat{\theta}_1$.

(ii)In general $LL(\theta_1)$ will be between two carefully chosen points X and Y, where initially X=0 and $Y=\hat{\theta}_1$.

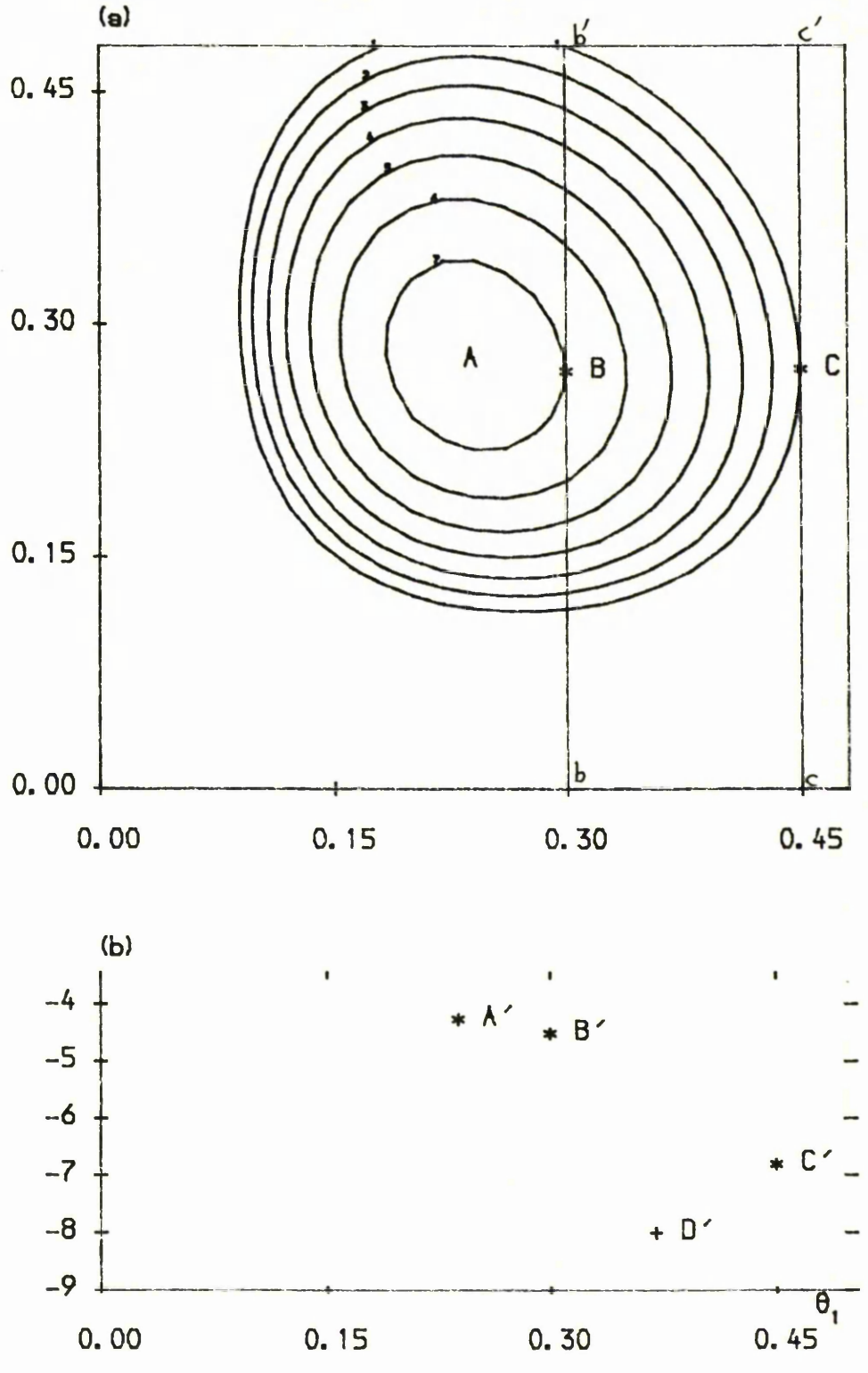
(iii)Using X and Y, find a suitable point Z nearer to the solution. Usually take Z as the root of the straight line passing between X and Y.

(iv)If $|G(Z)|<10^{-3}$, then $LL(\theta_1)=Z$; otherwise move to step (ii) but with different value for X and Y as described in the flow chart.

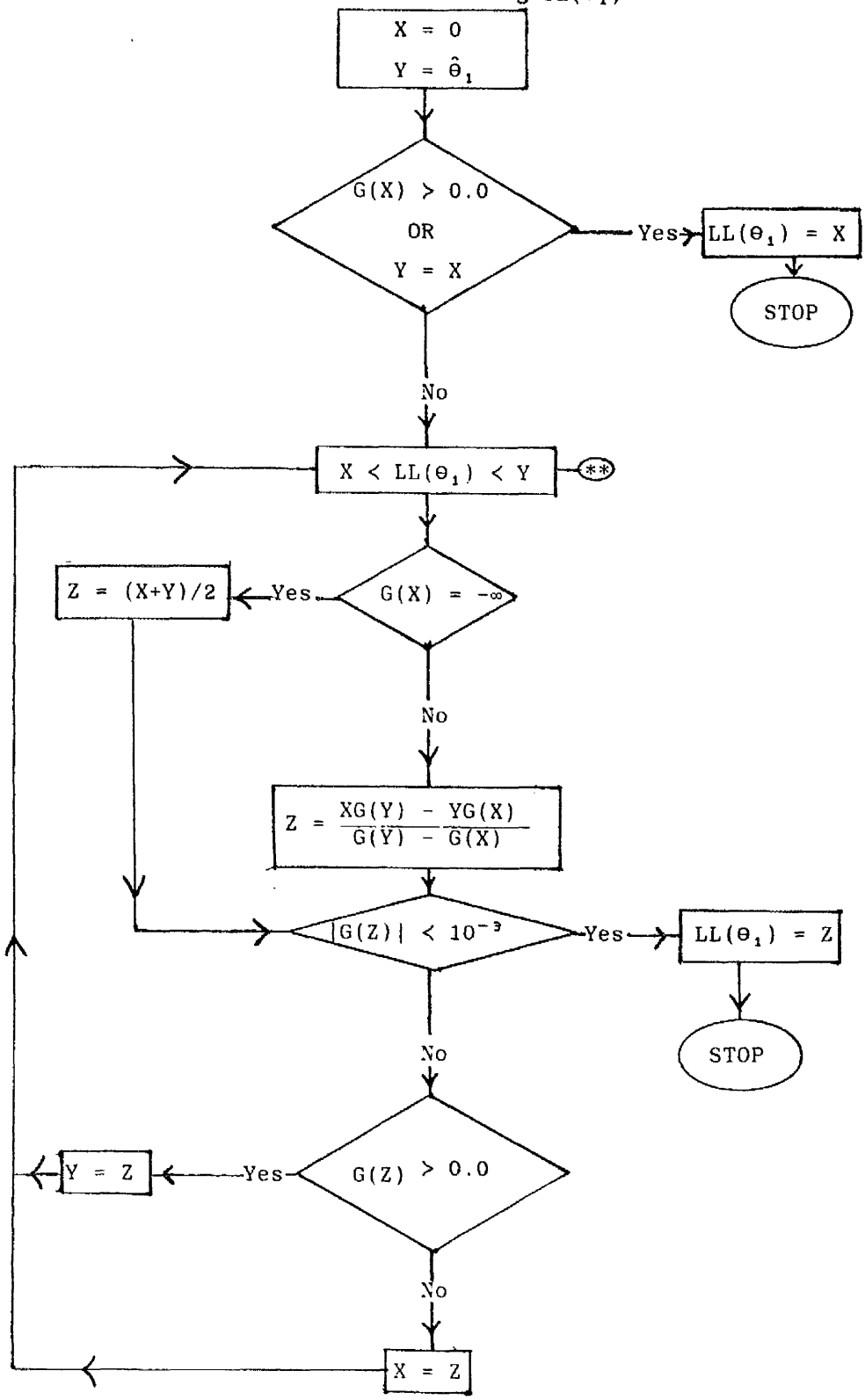
(Note that to find $UL(\theta_1)$ very similar argument will be adopted).

But to find G(Z) at any of the above stages we need to evaluate

FIGURE (4.4) A descriptive plot of the derivation of the profile ln likelihood when $r_1=1$ $r_2=5$ $r_3=6$ $r_4=13$



Figure(4.5) A flow chart for finding $LL(\theta_1)$



(To find $UL(\theta_1)$ the only change will be in the starting point,which will be X=0.5 and Y=$\hat{\theta}_1$, also at the step marked (**) the inequality signs will be the other way round).

$l(Z\ \theta_2^*(Z))$ numerically. Figure(4.6,a) is a flow chart showing the essential steps needed to find $l(a\ \theta_2^*(a))$ when Z=a. A brief description of it is as follows:

(i)Divide the range of θ_2 into eleven equally spaced points denoted by $\theta_2^{(j)}$, for j=0,1,..,10 and with $\theta_2^{(0)}$=0.0 and $\theta_2^{(10)}$=0.5. At each point calculate l_j, where $l_j = l(a\ \theta_2^{(j)}(a))$.

(ii)If at a special point j, $l_j > l_{j-1}$ and $l_j > l_{j+1}$, then θ_2^* is in the range defined by $[\theta_2^{(j-1)}, \theta_2^{(j+1)}]$. Now let $\theta_2^{(j-1)}$=X $\theta_2^{(j)}$=Z and $\theta_2^{(j+1)}$=Y.

(iii)$\theta_2^* \in R^{(i)} = [X,Y]$. Fit a quadratic to the three points X, Y and Z and their corresponding ln likelihoods. The location of the maximum of this quadratic, V, is a nearer approximation to θ_2^*.

(iv)If the required accuracy at the maximum is met, then stop. If not, a smaller range $R^{(i+1)}$ will be defined as shown in the chart. Move to (iii) to continue.

(v)Special care has to be taken at the boundaries, in case θ_2^*=0.0 or θ_2^*=0.5, this case is shown in the supplementary flow chart shown in figure(4.6,b).

(4.6.3)Ott and LSM approximations

Both methods depend on the quadratic approximation to the ln likelihood. Let:

$$l(\underline{\theta}) = l(\underline{\phi}) = Ql(\underline{\phi}) + \underline{e}$$

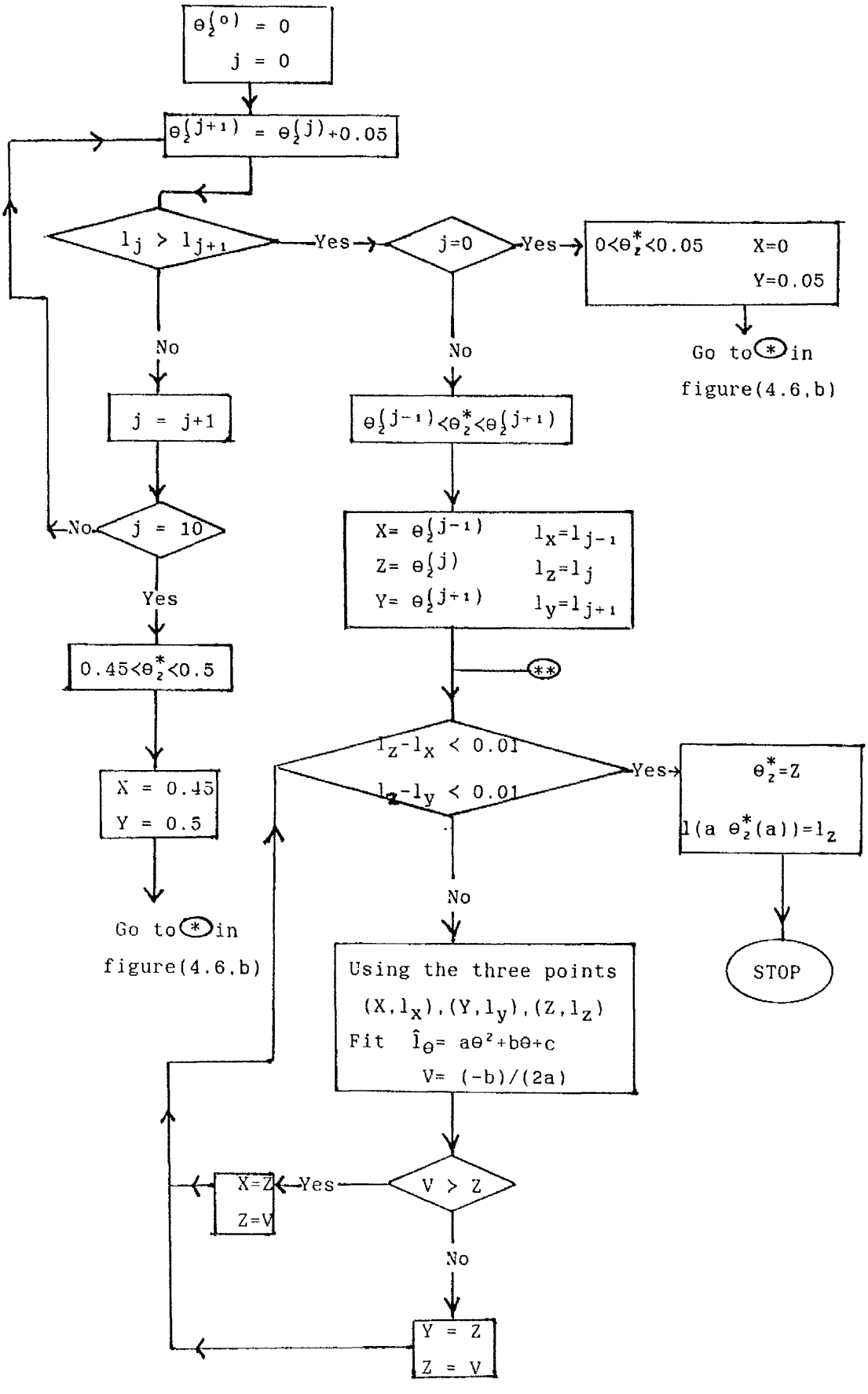
where $Ql(\underline{\phi})$ is the quadratic ln likelihood for either Ott or LSM approximations and $\underline{e} = l(\underline{\phi}) - Ql(\underline{\phi})$.

The unknown coefficients in (4.5) are determined by fitting $Ql(\underline{\phi})$ to some chosen points $(\underline{\phi}, l(\underline{\phi}))$, where $l(\underline{\phi})$ is the value of the original ln likelihood at the point $\underline{\phi}$.

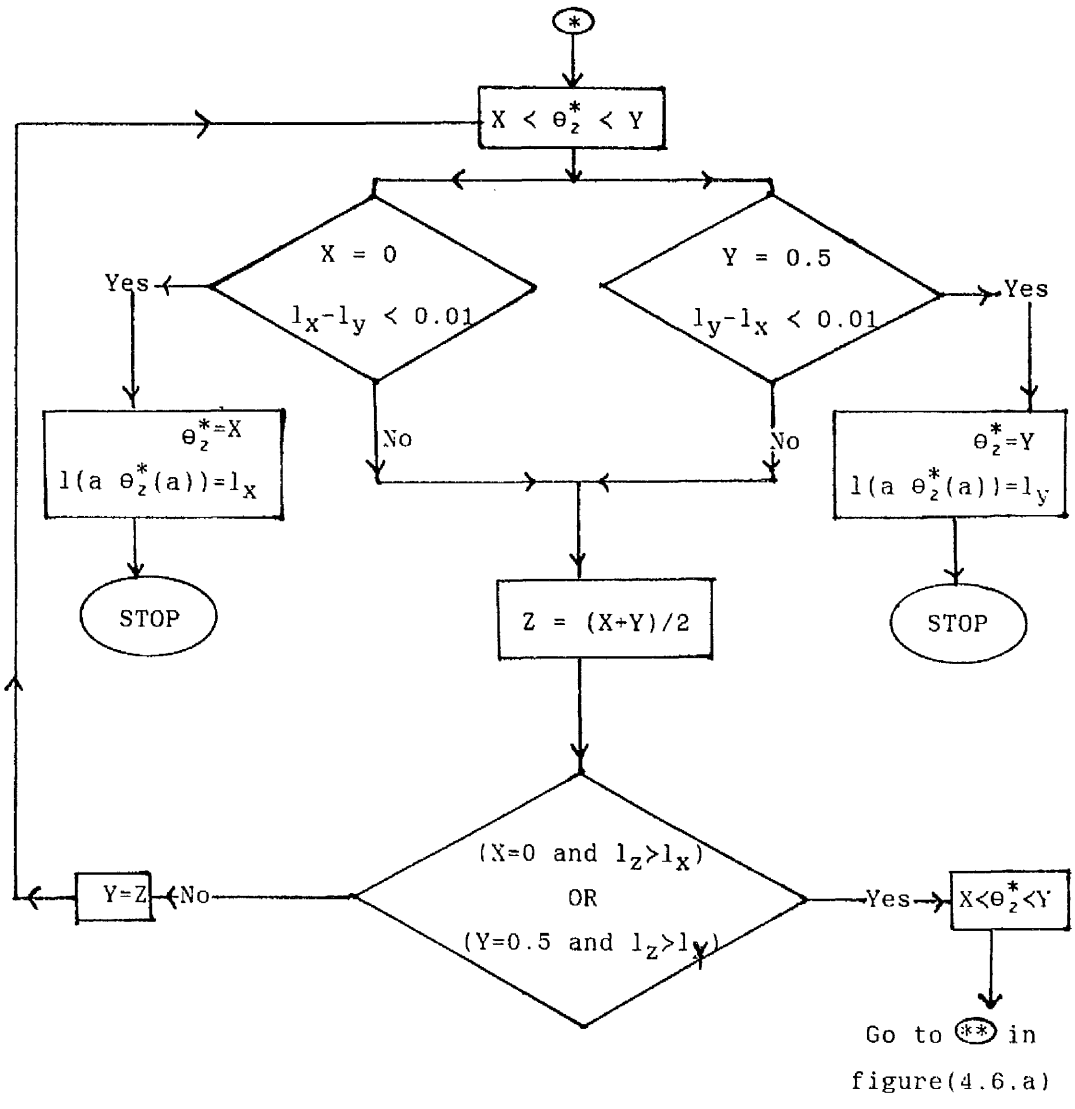
(A)The choice of points under Ott:

As seen before, six points near the maximum are needed. These points have been chosen as follows

Figure(4.6.a) Flow chart for finding $l(a\ \theta_2^*(a))$



Figure(4.6,b) continuation of figure(4.6,a)



(i)Evaluate the original ln likelihood ,$l(\underline{\theta})$ at $(\theta_{1i},\theta_{2j})$, for i=1,2..6 and j=1,2..6, where $\theta_{kl+1}= \theta_{kl}+0.1$ and k=1,2 , l=2,..5 and $\theta_{k1}=0.0$. From these 36 points, choose the first approximation to the maximum, denoted here by $(\theta_{1m},\theta_{2m})$.

(ii)Evaluate the original ln likelihood, $l(\underline{\theta})$ at $(\theta_{1i},\theta_{2j})$, for i=1,2..5 and j=1,2..5 where $\theta_{kl+1}= \theta_{kl}+0.05$ and k=1,2 , $\theta_{11}=\theta_{1m}-0.1$ and $\theta_{21}=\theta_{2m}-0.1$. (Note that some of these points would have been evaluated in (i)). From these 25 points, choose the second and last approximation to the maximum $(\theta_{1M},\theta_{2M})$ plus five other points around it arranged as in table(4.4)(a), (b) or (c) according to the circumstances mentioned in the tables. Now let

$$\underline{L} = \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_6 \end{bmatrix} \qquad A = \begin{bmatrix} \phi_{11}^2 & \phi_{21}^2 & \dots & 1 \\ \phi_{12}^2 & \phi_{22}^2 & \dots & 1 \\ \vdots & & & \\ \phi_{16}^2 & \phi_{26}^2 & \dots & 1 \end{bmatrix} \qquad \underline{b} = [a_1\ a_2\ b_1\ b_2\ c\ d]^T$$

where (ϕ_{1i},ϕ_{2i}) are the square root transformation of the six chosen points $(\theta_{1i},\theta_{2i})$, for i=1,2..6. Thus, $\underline{L} = A\underline{b} + \underline{e}$ and a unique and perfect solution, (i.e) with $\underline{e} = \underline{0}$, of the unknown vector $\underline{b}$ would be:

$\underline{b} = A^{-1}\underline{L}$, given that A is nonsingular.

(B)The choice of points under LSM:

The concept of the LSM, is to determine the quadratic likelihood, $QlL(\underline{\phi})$, by using points near an approximate 95% contour of the likelihood so that the final interval will be near the original one; to do so, we decided to choose 4, 8 and 4 points on the 92.5%, 95% and the 97.5% Ott's contours respectively. This has been done as follows:

(i)Using a suitable reparametization, the 95% ellipse, for example, could be transformed to a circle. Using formula (4.6,a),

Table(4.4) Choice of points under the Ott approximation

(a)If the maximum(θ_{1M} θ_{2M}) does not occur at any of the boundaries

θ_2 \ θ_1	θ_{1M-1}	θ_{1M}	θ_{1M+1}
θ_{2M-1}		l_1	
θ_{2M}	l_2	l_3	l_4
θ_{2M+1}		l_5	l_6

(b)If the maximum occurs at one of the boundaries

θ_2 \ θ_1	0.0	0.05	0.10
θ_{2M-1}	l_1	l_2	l_3
θ_{2M}	l_4	l_5	
θ_{2M+1}	l_6		

(Similar tables could be produced if (θ_{1M} θ_{2M}) = (0.5 θ_{2M}) or (θ_{1M} 0.0) or (θ_{1M} 0.5)).

(c)If the maximum occurs at two of the boundaries

θ_2 \ θ_1	0.0	0.05	0.10
0.0	l_1	l_2	l_3
0.05	l_4	l_5	
0.10	l_6		

(Similar tables could be produced if (θ_{1M} θ_{2M}) = (0.0 0.5) or (0.5 0.0) or (0.5 0.5)).

the equation of this ellipse would be defined as follows:

$$Q1O(\underline{\Phi}) - Q1O(\hat{\underline{\Phi}}_O) = -(0.5)x^2(2;0.95) \quad \leftrightarrow$$

$$(\underline{\Phi} - \hat{\underline{\Phi}}_u)^T K (\underline{\Phi} - \hat{\underline{\Phi}}_u) = x_1^2(2;0.95) - 2Q1(\hat{\underline{\Phi}}_O) + 2Q1(\hat{\underline{\Phi}}_u)$$

$$= \text{constant} \qquad (4.12)$$

where $\hat{\underline{\Phi}}_u$ is the unbounded maximum of $Q1(\underline{\Phi})$ and

$$K = \begin{bmatrix} -2a_1 & -c \\ -c & -2a_2 \end{bmatrix}$$

If K is positive definite then,

$$K = Q \Lambda Q^T \quad , \text{ where } \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad ; \quad Q = [\underline{u}_1 \;\; \underline{u}_2]$$

and λ_1, λ_2 are the positive eigenvalues of K and $\underline{u}_1$, $\underline{u}_2$ are the corresponding normalised eigenvectors. It follows then

$$K = \lambda_1 \underline{u}_1 \underline{u}_1^T + \lambda_2 \underline{u}_2 \underline{u}_2^T$$

Now let

$$\Psi_1 = \sqrt{\lambda_1}\; \underline{u}_1^T(\underline{\Phi} - \hat{\underline{\Phi}}_u)$$

$$\Psi_2 = \sqrt{\lambda_2}\; \underline{u}_2^T(\underline{\Phi} - \hat{\underline{\Phi}}_u)$$

Therefore (4.12) will be equivalent to

$$\Psi_1^2 + \Psi_2^2 = \text{constant} \qquad (4.13)$$

The ellipse defined in (4.12) has been transformed to a circle with centre lying at the origin (0,0) and with radius equal $\sqrt{\text{constant}}$, when using the new parameters Ψ_1 and Ψ_2.

(ii)Using the same reparametrization, transform the feasible region in the domain of $\underline{\Phi}$ to the domain of $\underline{\Psi}$, where $\underline{\Psi} = [\Psi_1 \; \Psi_2]^T$:

$$0.0 < \Phi_1 < \sqrt{0.5} \qquad\qquad 0.0 < k_1\Psi_1 + k_2\Psi_2 + k_3 < \sqrt{0.5}$$

$$\leftrightarrow$$

$$0.0 < \Phi_2 < \sqrt{0.5} \qquad\qquad 0.0 < k_4\Psi_1 + k_5\Psi_2 + k_6 < \sqrt{0.5}$$

where $k_1, k_2, \ldots, k_6$ are suitable constants that make the above transformation true.

(iii)Choose eight equally spaced points on the circle (4.13), but within the feasible region. If the whole circle lies within the

feasible region, then the eight chosen points will be:

Ψ_{1i} = √constant * sin((i-1)π/4)

Ψ_{2i} = √constant * cos((i-1)π/4) for i=1,2,..8.

(iv)In a similar manner choose the other eight points on the 92.5% and 97.5% ellipses. Transform back all the sixteen points to the $\underline{\Phi}$ domain. Use them to determine the unknown coefficients of Q1($\underline{\Phi}$), (i.e) let:

$$\underline{L} = \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_{16} \end{bmatrix} \qquad A = \begin{bmatrix} \Phi_{11}^2 & \Phi_{21}^2 & \dots & 1 \\ \Phi_{12}^2 & \Phi_{22}^2 & \dots & 1 \\ \vdots & & & \\ \Phi_{1,16}^2 & \Phi_{2,16}^2 & \dots & 1 \end{bmatrix} \qquad \underline{b} = [a_1\ a_2\ b_1\ b_2\ c\ d]^T$$

So that

$\underline{L} = A\underline{b} + \underline{e}$

Now using least square methods:

$\hat{\underline{b}} = (A^TA)^{-1}A^T\underline{L}$

(C)Joint IE for Θ:

To completely determine the joint interval defined in (4.6,a),(4.6,c) for the Ott and the LSM approximation respectively, all we need is to find $\hat{\underline{\Phi}}_O$, (i.e) the MLE of Q1O($\underline{\Phi}$). This is just a maximization problem subject to bounds on the variables, see appendix(4.2).

(D)Marginal IE for Θ_i:

The marginal interval for Θ_1 defined in (4.7,b) will be completely determined when $\Phi_2^*(\Phi_1)$ is evaluated. From appendix(4.2), it has been found that $\Phi_2^*(\Phi_1)$ is one of three possible linear functions of Φ_1. Therefore Q1(Φ_1 $\Phi_2^*(\Phi_1)$) is the combination of one or at most three different quadratics in Φ_1 within the feasible range [0.0 √0.5]. Therefore the set of points of Θ_1 that satisfy (4.7,b) are all the values of Θ_1 between the squares of the lower and upper roots of the following function:

$$G(\phi_1) = Ql(\phi_1\ \phi_2^*(\phi_1)) - Ql(\hat{\underline{\phi}}_0) + 1.92$$

provided that the roots are within the feasible range. Special care has to be taken when finding the roots of $G(\phi_1)$ to account for the three possible forms of $Ql(\phi_1\ \phi_2^*(\phi_1))$.

(4.7)Results, phase known situation

In table(4.5), the exact CC and EL are given under the different combinations of methods and parameters, the following are some observations concerning these results:

(a)The last column in the table which gives the total probability used in deriving the CC and El, ((i.e) $\Sigma_r P_r$ where $r \epsilon R_s$), shows a satisfying value ≥ 0.995 for all the different values of $\underline{\theta}_t$.

(b)The joint CC under the different methods are very near the 0.95 threshold apart from the odd situation of 0.91 when using the Ott approximation and with θ_{1t}=0.1 and θ_{2t}=0.15. On the other hand, when comparing the CC of the approximate methods to the original likelihood, no obvious conclusion can be made, although we can say that perhaps the LSM is slightly nearer the original likelihood method.

(c)When comparing the marginal CC of the three methods,the first thing we noticed was that there is nearly no interaction between the two recombination fractions θ_{1t} and θ_{2t}. The CC and EL of θ_{it} seems to be the same whatever value of θ_{jt} is used. This last comment had led us to suggest producing the same kind of results when N=25, θ_{1t}=0.05,0.1 or 0.15 and θ_{2t} is varying from 0.01 to 0.25 in step of 0.01, and thus be able to produce figures showing the variation of the joint CC, marginal CC and EL against θ_{2t} for each θ_{1t}; figures (4.7), (4.8) and (4.9) respectively, see next paragraph. Also, we can see that the Ott approximation gave two extreme results, a very high CC of 0.99 associated with a shorter EL when θ_{1t} or θ_{2t}=0.05, and a low CC of 0.88 when θ_{1t} or

Table(4.5) Confidence coefficient and expected length of the phase known situation, and for the various combination of θ_t

θ_{1t}		Joint	Marginal for θ_1		Marginal for θ_2		Total
θ_{2t}	Method	CC	CC	EL	CC	EL	P_r
0.05	Org.	0.96	0.97	0.16	0.97	0.16	
0.05	Ott	0.96	0.99	0.15	0.99	0.15	0.999
	LSM	0.96	0.97	0.16	0.97	0.16	
0.05	Org.	0.94	0.97	0.16	0.90	0.22	
0.10	Ott	0.94	0.99	0.15	0.91	0.21	0.999
	LSM	0.94	0.97	0.16	0.90	0.22	
0.05	Org.	0.94	0.97	0.16	0.95	0.26	
0.15	Ott	0.94	0.99	0.15	0.88	0.26	0.998
	LSM	0.94	0.97	0.16	0.96	0.26	
0.10	Org.	0.93	0.90	0.22	0.90	0.22	
0.10	Ott	0.94	0.91	0.21	0.91	0.21	0.998
	LSM	0.94	0.91	0.22	0.91	0.22	
0.10	Org.	0.93	0.90	0.22	0.95	0.26	
0.15	Ott	0.91	0.93	0.21	0.88	0.26	0.997
	LSM	0.93	0.91	0.22	0.96	0.26	
0.15	Org.	0.94	0.95	0.26	0.95	0.26	
0.15	Ott	0.92	0.88	0.26	0.88	0.26	0.995
	LSM	0.94	0.96	0.26	0.96	0.26	

θ_{2t}=0.15. On the other hand, althought the LSM's CC is sometimes less then the Ott's CC, it seems to be more stable along the different values of $\underline{\theta}_t$ varying from 0.90 to 0.97, as well as being nearer the original CC.

From figure(4.7) where (a), (b) and (c) correspond to θ_{1t}=0.05, 0.1 and 0.15 respectively, we can safely say that, firstly, the LSM curve is nearer the original one, and secondly, for θ_{1t}=0.1 or 0.15, all the three methods seem to be somewhat short of the 0.95 threshold, although both the original and LSM seem to give a better result as compared to the Ott approximation. From figure(4.8)(a,c,e), which are the marginal CC of θ_1 against θ_{2t} for θ_{1t}=0.05, 0.1 and 0.15 respectively, we can still see that there is no or a very slight interaction between the two recombination fractions, this is shown in somewhat straight line curves for all the three methods, apart from the LSM for θ_{1t}=0.05. Also the three plots seem to suggest that the result of the LSM is usually between that of Ott and the original likelihood but nearer the latter method. As for figure(4.8)(b,d,f) which are the marginal CC of θ_2 against θ_{2t} for each θ_{1t}, we can see again, and because of the lack of interaction between the θ's, that the three plots are nearly the same. Nevertheless, we can see that, although the LSM is usually nearer the original likelihood method, it seems to give the most stable and highest CC's results, apart from θ_2<0.07 where the Ott approximation has the highest CC for this range of θ_2, but where also all the three methods give a result way above the 0.95 threshold. The slightly higher CC of LSM is reflecting itself, as would be expected, in a slightly longer EL of the three methods, as shown in figure(4.9) for both θ_1 and θ_2 and under the various values of θ_{1t}. What is somehow unexpected is the high CC of the

FIGURE (4.7) The joint confidence coefficient against θ_2, for the original likelihood, Ott & LSM

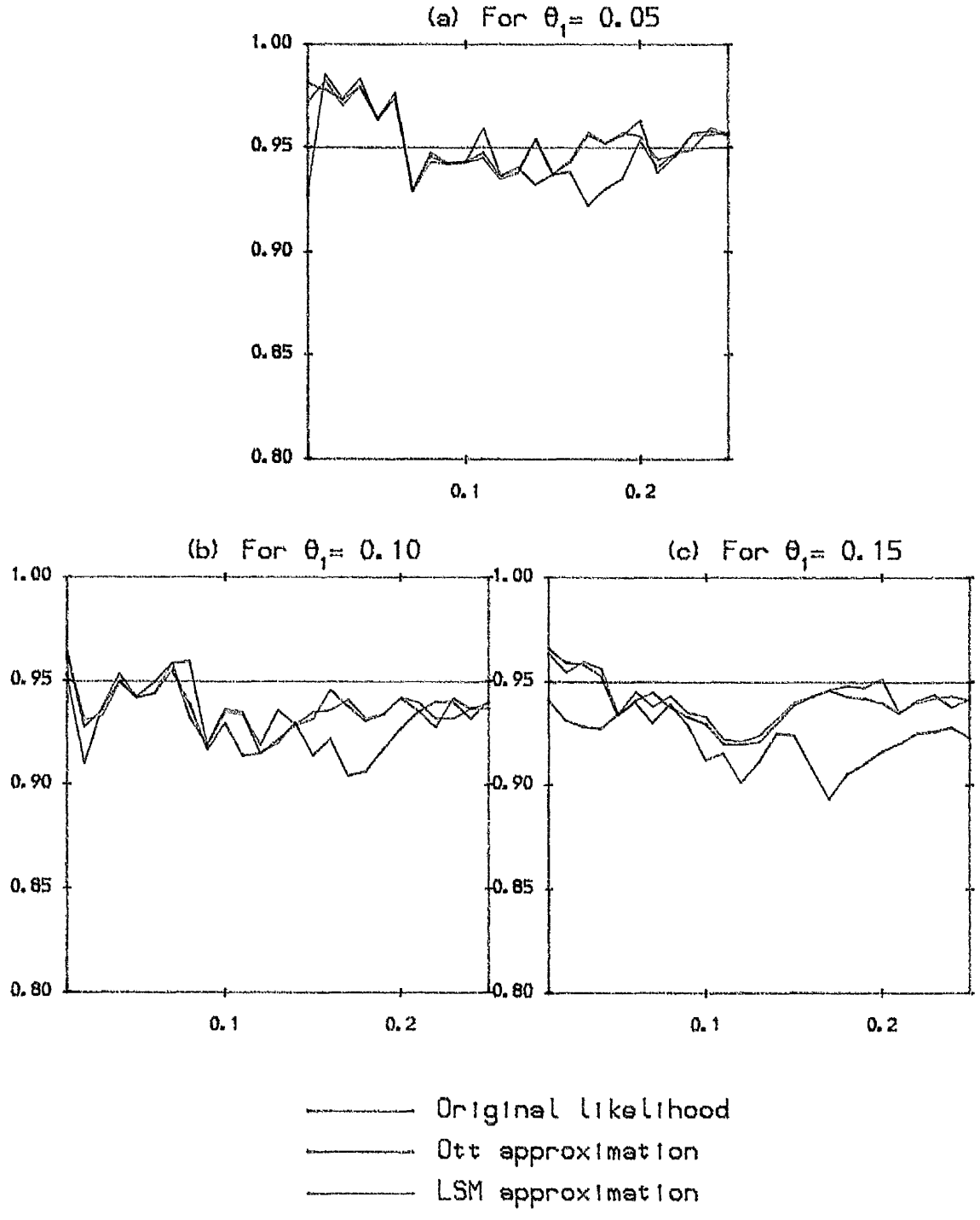


FIGURE (4.8) The marginal confidence coefficient against $\theta_{2'}$ for the original likelihood, Ott & LSM

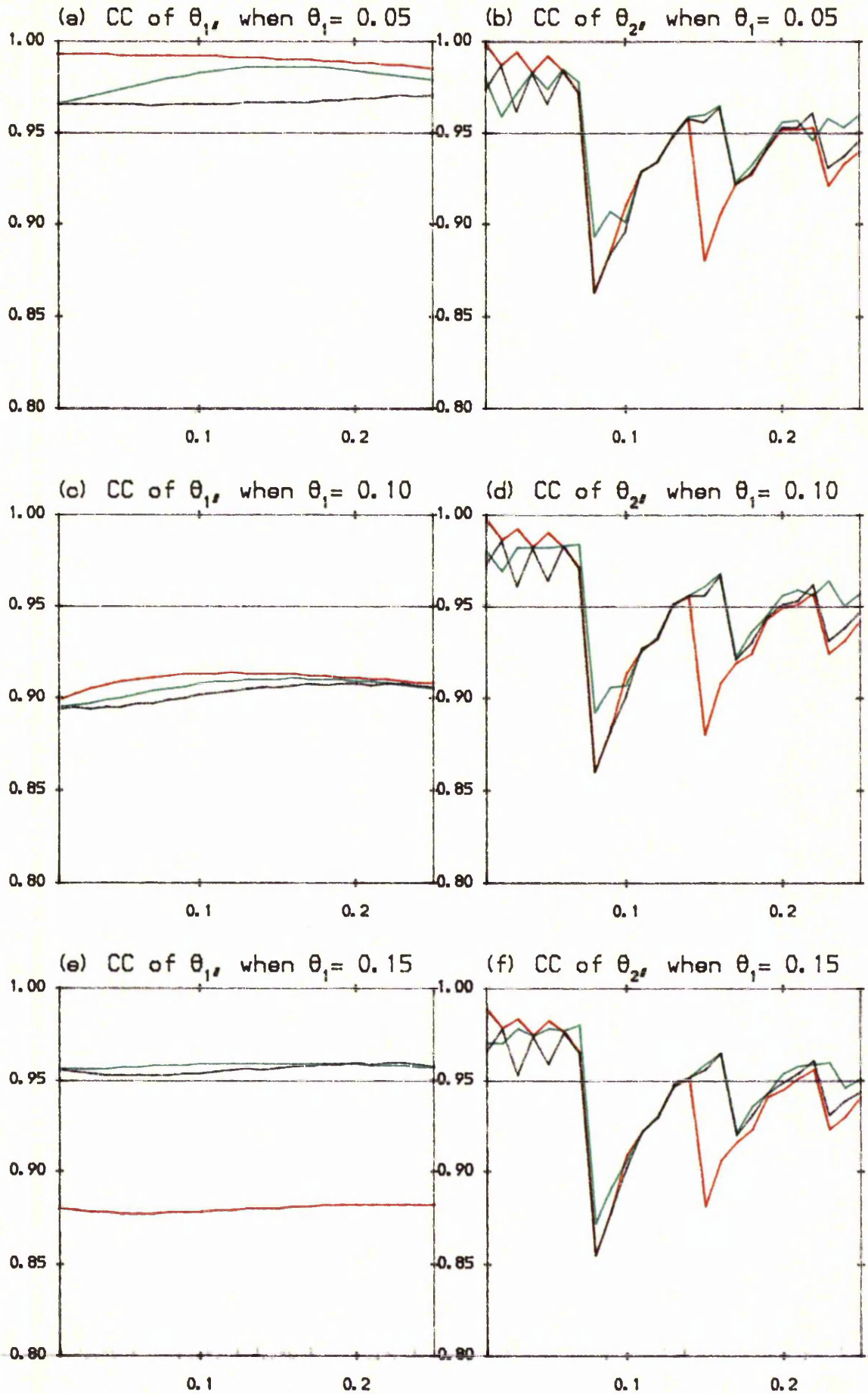
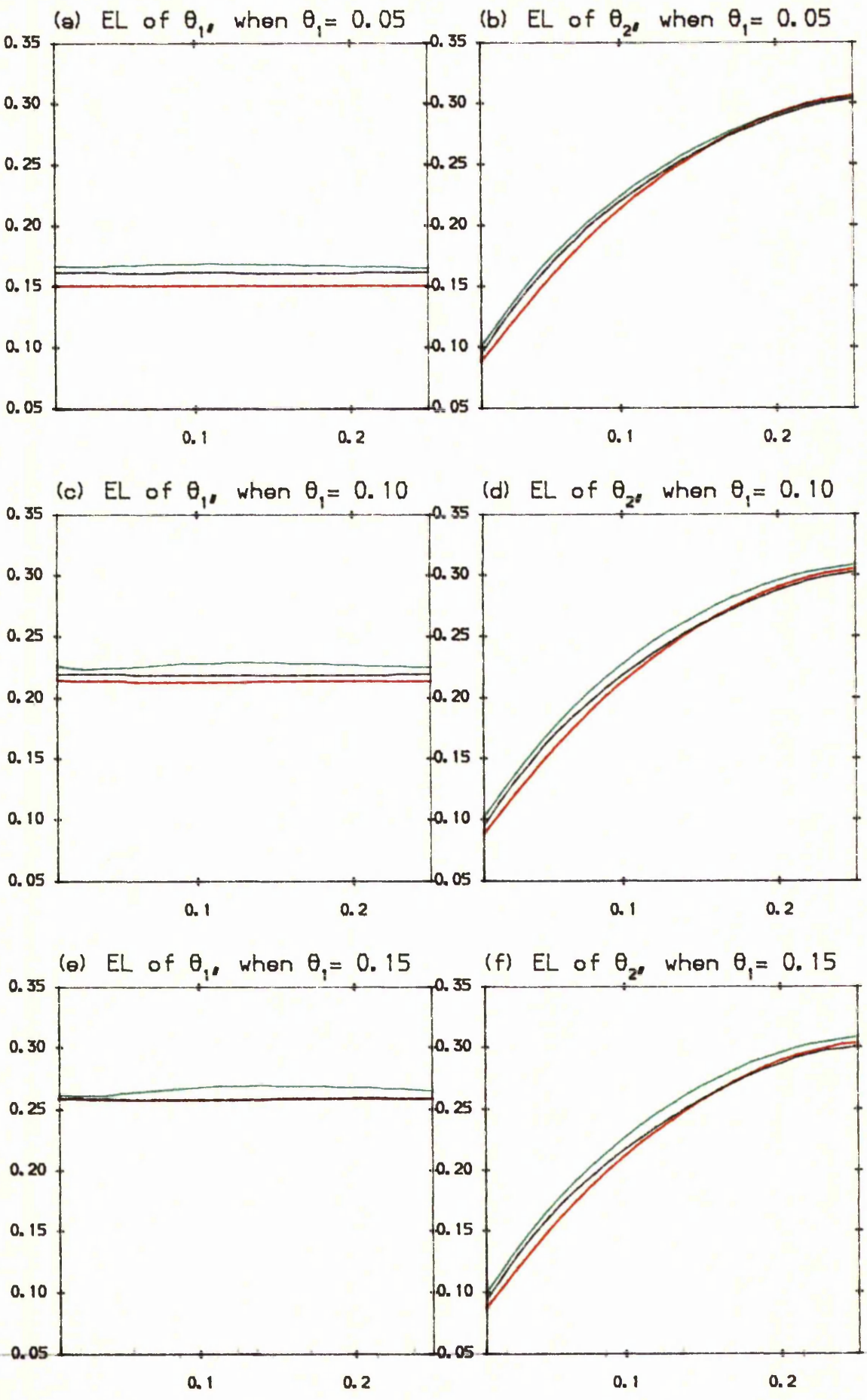


FIGURE (4.9) The expected length against θ_2, for Ott & LSM



Ott method for $\theta_{2t}<0.07$ although it has, in general the shortest EL. One explanation would be that the majority of the data points, $\underline{r}$, would lead to a shorter marginal interval when using Ott except for some odd ones which would lead to a longer one. Actually, a high proportion of the data points will lead to a fairly nearly quadratic $l(\underline{\phi})$, and therefore will lead to very close 95% ellipses when using all the three methods; other points which will lead to an unquadratic $l(\underline{\phi})$, will bring out to the light the difference between the three methods, some of these later points are used to plot the three 95% joint regions for $\underline{\phi}$, they are shown in figure(4.10).

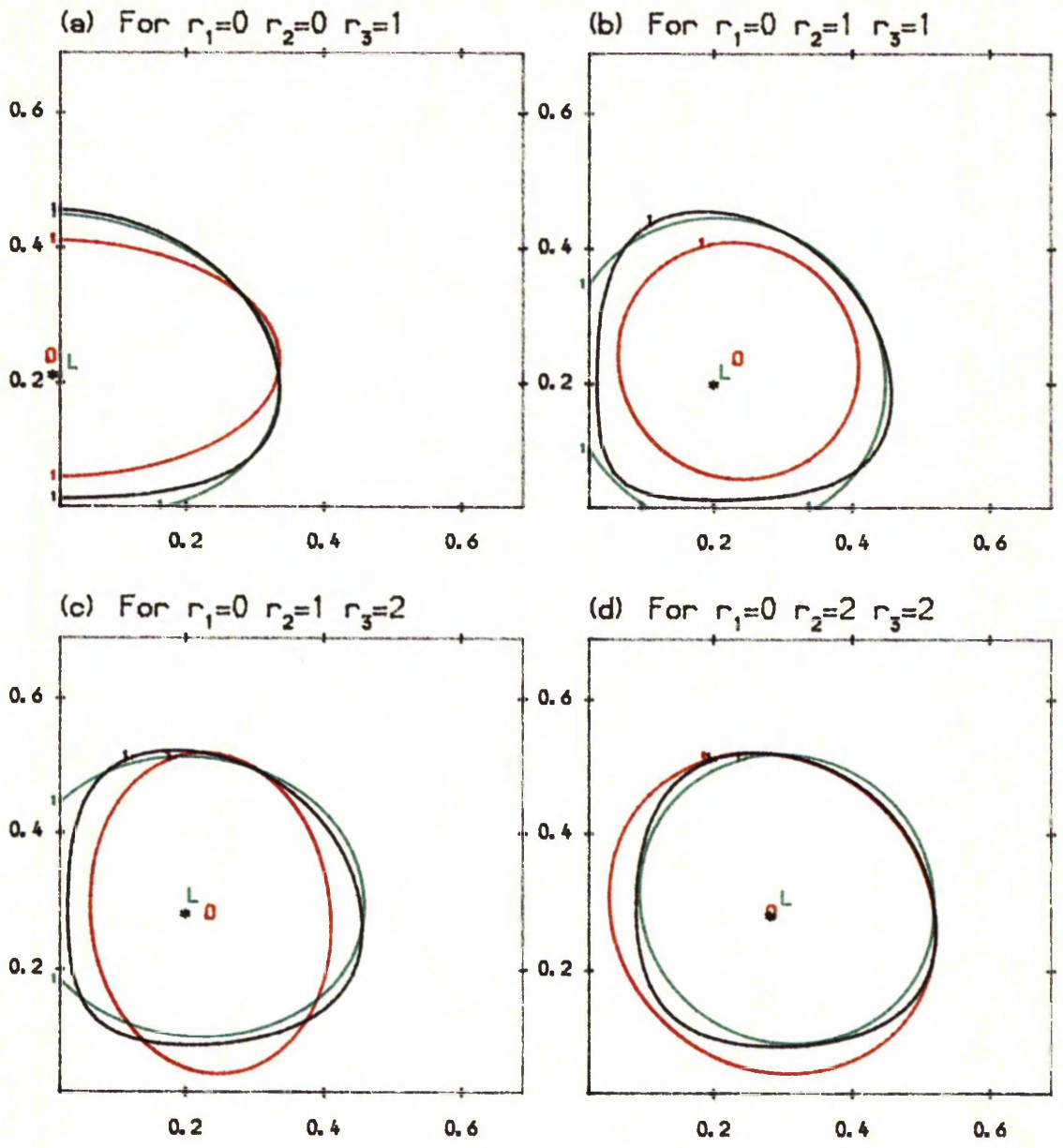
The shortness of the marginal interval when using Ott, is again reflected very clearly in the histogram plots of the distribution of the difference in the marginal interval length between Ott and the LSM methods, denoted previously by D(m). But, because of the lack of interaction between the two θ's, we decided to show only the histogram plots of D(1), (i.e) for a marginal interval for θ_1, when $\theta_{1t}=\theta_{2t}=0.05$, 0.1 or 0.15, which are shown in figure(4.11) (a), (b) and (c) respectively. These figures, suggest strongly that the probability of having a shorter marginal interval for θ, when the Ott approximation is used, is always greater than 0.5, and that this probability becomes smaller as θ becomes larger.

(4.7) Application, given the phase unknown situation

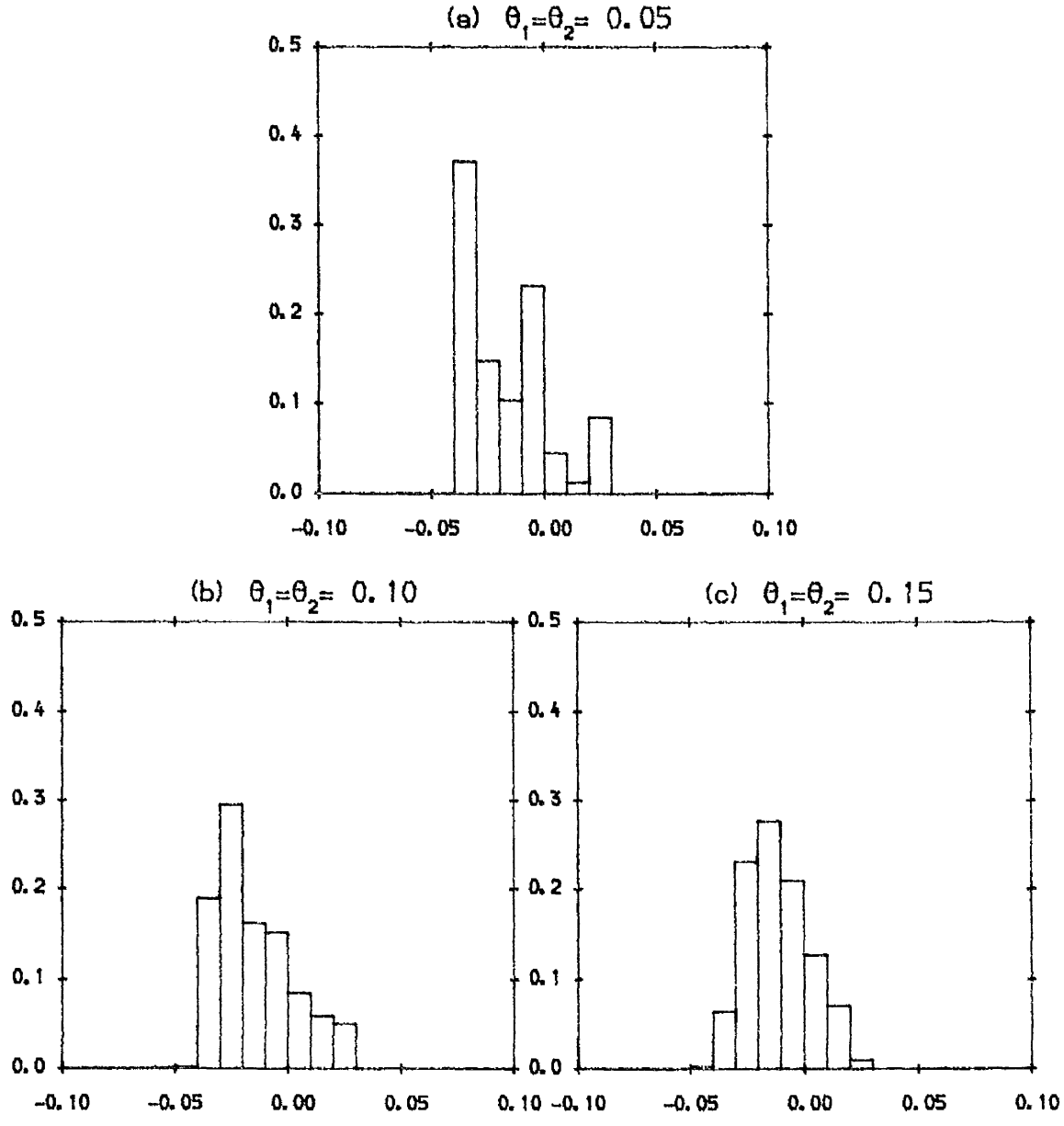
This section discusses the application of the three methods of IE discussed previously, when using a data vector $\underline{r}$ arising from the multinomial distribution of (1.15).

In figure(4.12)(a,b,c,d), the contour plots of $l(\underline{\theta})$ for some data vector $\underline{r}$ are shown, from which we can see that $l(\underline{\theta})$ is neither unimodal nor concave. Although the recipe of both the

FIGURE (4.10) Some 95% joint region for $(\phi_1\ \phi_2)$ using the original likelihood, Ott & LSM



FIGURE(4.11) Histogram plots of the difference in length between Ott & LSM's interval for θ_1

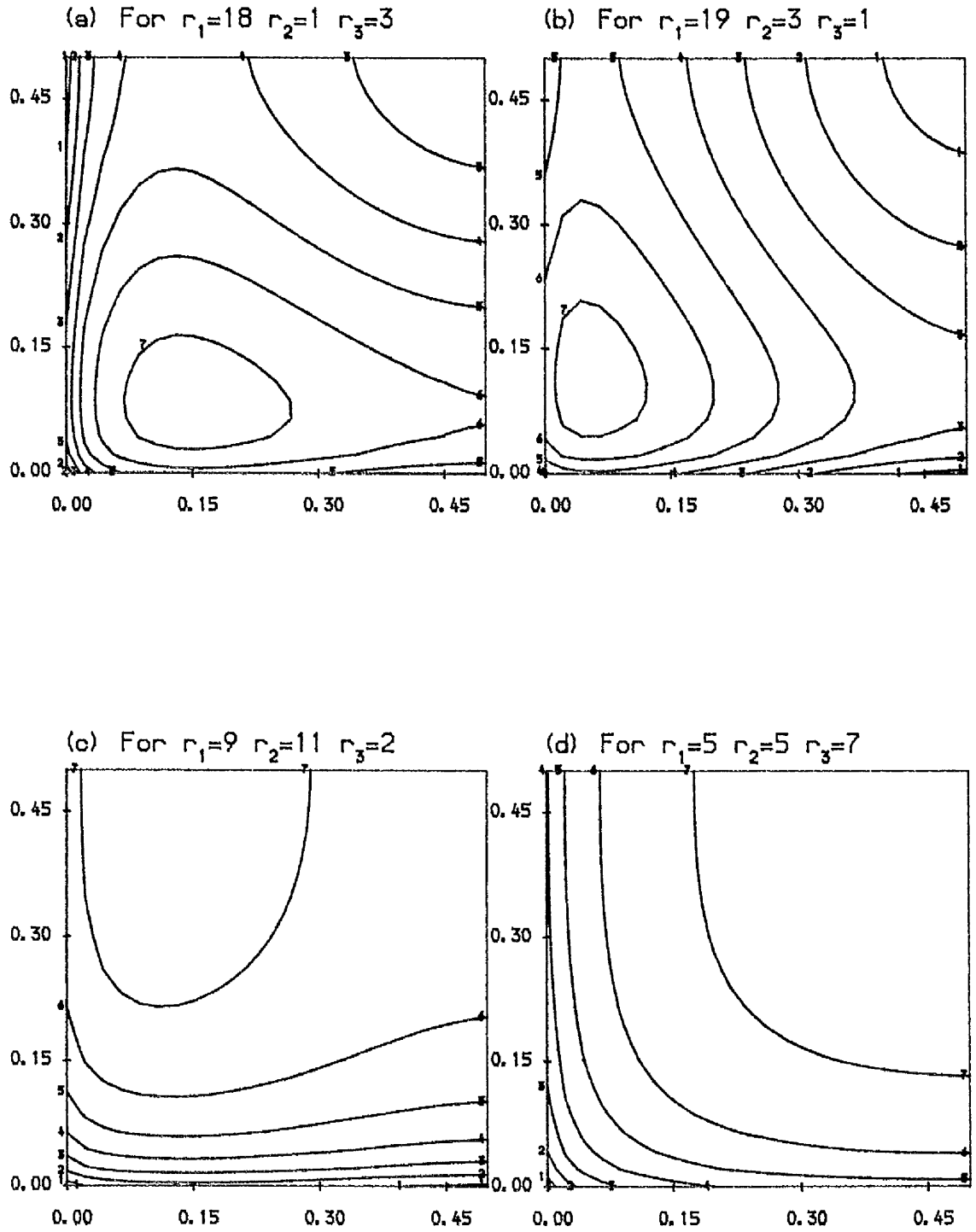


joint and marginal intervals will still be applicable, we will find that using any of the prescribed numerical techniques to calculate the intervals would be dangerous. But in order to help ourself in understanding the behaviour of the likelihood function for this situation, we decided to compare the likelihood to its corresponding one if the Haldane map function is assumed instead. In figure(4.13), the contour plot of the likelihoods given either map function are shown for the same data point $\underline{r}$ as before, which seem to suggest that both likelihoods are fairly alike. An analytic maximisation of $l(\underline{\theta})$ given the Haldane map function is a straight forward procedure (see appendix (4.3)), from which we can see that the likelihood will have one maximum within the feasible region if $r_3+r_4<(n/2)$ and $r_2+r_4<(n/2)$. If either or both conditions are not satisfied then the likelihood will have a unique maximum occuring at at least one of the following boundaries, $\theta_1=0.5$ or $\theta_2=0.5$.

Applying either Ott of the LSM to this kind of likelihood, even when the Haldane map function is assumed will not be realistic, as it seems quite clear from the ln likelihood contour plots that the ln likelihood is far away from being quadratic. Nevertheless if we can find a one-to-one transformation to $\underline{\theta}$, (eg) $\underline{\eta}=g(\underline{\theta})$, such that $l(\underline{\eta})$ is approximately quadratic or just concave and unimodal, then applying either Ott or the LSM methods for finding the required intervals for $\underline{\eta}$ will be a straight forward procedure. A 95% IE for $\underline{\theta}$ or θ_i will be just the induced interval of $\underline{\eta}$ or η_i. We suspect that the transformation $\eta_i=2\theta_i(1-\theta_i)$ for i=1,2, which is a one-to-one transformation of $\underline{\theta}$ within the feasable region, will satisfy the above requirement.

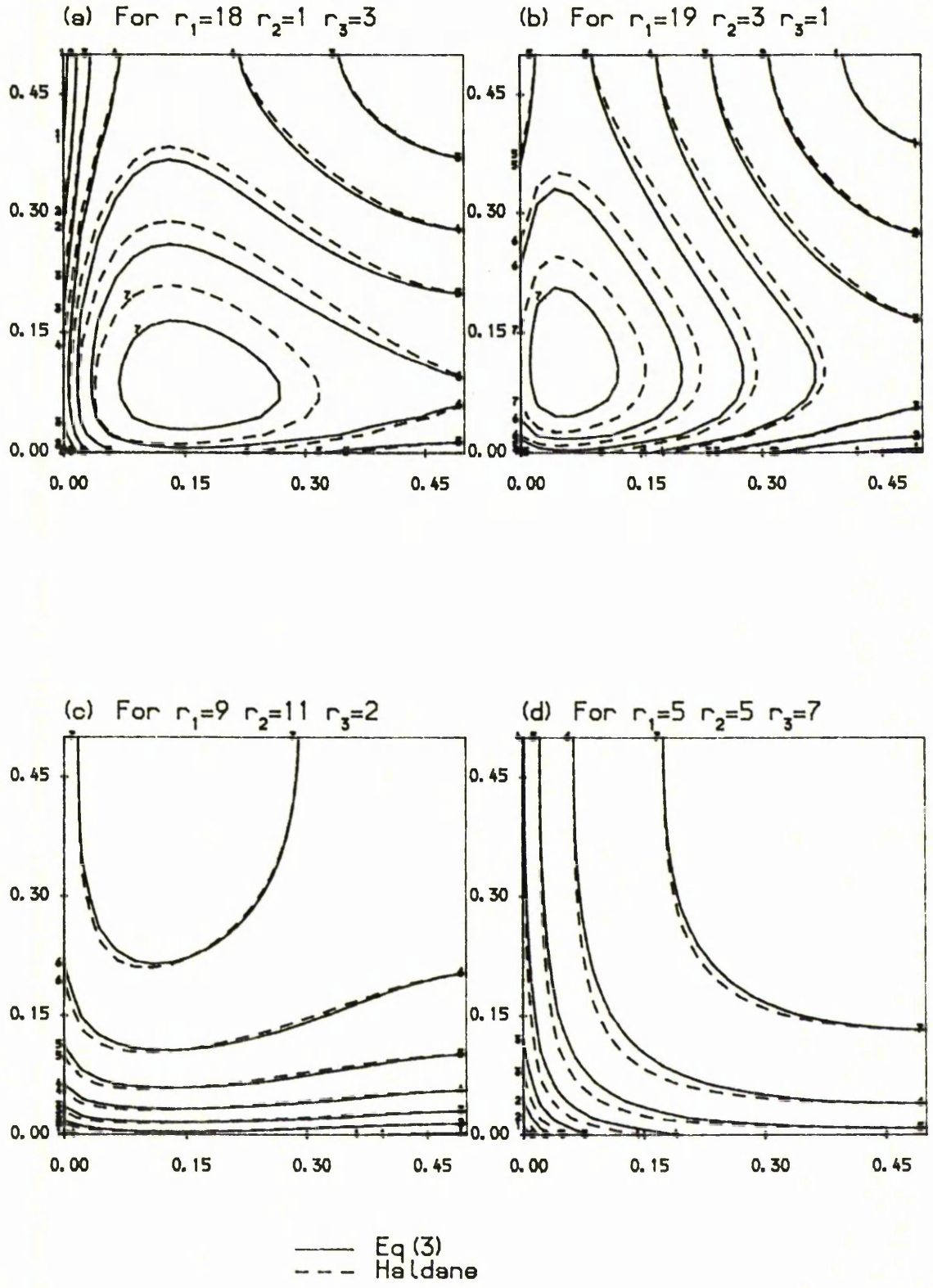
PHASE UNKNOWN SITUATION

FIGURE (4.12) Contour plot of the ln likelihood when N=25



PHASE UNKNOWN SITUATION

FIGURE (4.13) Contour plot of the ln likelihood when N=25 & given either Eq(3) or the Haldane map function



CHAPTER FIVE: Comparing intervals using the likelihood and the Bayesian approaches

(5.1) Introduction

The aim of this chapter is to compare the performance of interval estimates constructed using the likelihood approach as described in chapter 4, when the original likelihood was used, and the Bayesian approach.

Methods and results are produced for the simplest model in genetics of two loci situation and either a phase known or unknown double backcross mating, whereas only method is mentioned for the three loci situation.

(5.2) Bayesian approach

(5.2.1)Introduction

In general, if $\underline{\theta}$ is the parameter vector of interest with sample space Θ, then using the Bayesian approach means that $\underline{\theta}$ is essentially regarded as a random variable. Therefore any knowledge we have about the true value of $\underline{\theta}$, at any stage, can be expressed by a probability distribution over Θ.

Let $\pi(\underline{\theta})$ be the prior probability distribution of $\underline{\theta}$ and $p(\underline{r}|\underline{\theta})$ be the density function of the sample data vector $\underline{r}$ specified by the probability model, then the posterior probability function of $\underline{\theta}$, $\pi(\underline{\theta}|\underline{r})$ will be:

$$\left.\begin{aligned} &\pi(\underline{\theta}|\underline{r}) = \text{const}\ \pi(\underline{\theta})\ p(\underline{r}|\underline{\theta}) \\ &\text{where } (\text{const})^{-1} = \int_{\Theta} \pi(\underline{\theta})p(\underline{r}|\underline{\theta})d\underline{\theta} \end{aligned}\right\} \quad (5.1)$$

(5.2.2) Interval estimate

Although $\pi(\underline{\theta}|\underline{r})$ constitutes the complete inferential statement about $\underline{\theta}$, a $100(1-\alpha)\%$ Bayesian confidence region, $IE(1-\alpha)$, can be defined as follows:

$$\int_{IE(1-\alpha)} \pi(\underline{\theta}|\underline{r})d\underline{\theta} = (1-\alpha) \qquad (5.2,a)$$

Several regions, IE(1-α), can satisfy the above definition. A unique interval can be defined, if $\pi(\underline{\theta}|\underline{r})$ is not uniform, by ensuring that the region IE(1-α) defined in (5.2,a) should be such that the probability density of every point inside it is at least as large as that of any point outside it.

(i.e) For $\underline{\theta}_1 \in IE(1-\alpha)$ and $\underline{\theta}_2 \notin IE(1-\alpha)$

$$\pi(\underline{\theta}_1|\underline{r}) \geq \pi(\underline{\theta}_2|\underline{r}) \qquad (5.2,b)$$

Therefore, $IE(1-\alpha) = \{\underline{\theta} \text{ such that } \pi(\underline{\theta}|\underline{r})>c\}$

where c is chosen such that:

$$\int_{\pi(\underline{\theta}|\underline{r})>c} \pi(\underline{\theta}|\underline{r})d\underline{\theta} = 1-\alpha \qquad (5.3)$$

Such interval will be the the shortest among all Bayesian confidence region of (1-α) confidence coefficient and is known as the highest probability density, HPD, interval.

Other criteria can be set to determine a unique (1-α) confidence region. A central confidence interval, CCI, (ie) cutting off equal tail area probabilities defines a unique (1-α) confidence region in one dimensional problems as follows:

$IE(1-\alpha) = \{\theta \text{ such that } \theta \in [\theta_1, \theta_2]\}$

where θ_1 and θ_2 are chosen such that

$$\int_{-\infty}^{\theta_1} \pi(\theta|r)d\theta = \int_{\theta_2}^{\infty} \pi(\theta|r)d\theta = \frac{\alpha}{2} \qquad (5.4)$$

Both intervals are going to be used in this chapter.

(5.2.3) Prior distribution

Two prior distributions from the genetic literature can be used. The first one, introduced by Haldane and Smith(1947) is based on the assumption of a uniform distribution for the

recombination fraction between the two loci, (i.e):

$$\pi(\theta) = 2 \qquad 0<\theta<0.5 \qquad (5.5)$$

The second one, introduced by Renwick(1971) to find the prior probability of x, $\pi(x)$, where x is the map distance between two autosomal loci, could be adjusted to find $\pi(\theta)$, the prior distribution of the recombination fraction between two loci known to be on the same chromosome.

The following arguments were introduced by Renwick to find $\pi(x)$, where x is defined as above:

Let A_i, i=1,2,..22, be the length in Morgan, the unit which measures the map distance x, of chromosome i and $T=\Sigma_i Ai$ be the total autosomal length. Further assume that:

(a)The locus on the chromosome is treated as a point on a line.

(b)A chosen autosomal locus is equally likely to occur at any point on the autosomal complement of length T; (i.e) the distribution of its position is uniform[0,T].

(c)The two loci under consideration are selected at random from the total.

Under these assumptions, the probability that a locus is on chromosome i is (A_i/T) and the probability of synteny, (i.e) that both loci are on the same chromosome, is $\Sigma_i(A_i^2/T^2)$. Therefore the probability density of the map length x between two syntenic loci on the chromosome A_i will be:

$$\pi(x \cap \text{both on } A_i) = \frac{2(A_i-x)}{T^2} \qquad 0<x<A_i$$

Now, let $A_i>A_{i+1}$ for i=1,2,...,21, then:

$$\Pr\{\text{synteny}\} = \Pr\{0<x<A_1\} = \sum_1 \int_0^{A_i} \frac{2(A_i-x)}{T^2}\,dx$$

$$= \sum_1 \frac{A_i^2}{T^2}$$

If the two loci are asyntenic, (i.e) are not on the same chromosome, this means that the map distance is infinite. For convenience assume that x=1000 if the two loci are asyntenic, then:

$$\Pr\{\text{asyntenic}\} = \Pr\{x=1000\} = 1 - \Sigma_i \frac{A_i^2}{T^2}$$

Adjusting this argument in order to suit our assumption of the two loci being on the same chromosome of length A, for example, will lead to the following $\pi(x)$:

$$\pi(x) = \frac{2(A-x)}{A^2} \qquad 0<x<A \qquad (5.6)$$

As an example take A to represent the map length of the longest human chromosome, chromosome number 1, which is estimated to be slightly over 3 Morgans (according to Renwick 1971), actually this number refer to the male-female average map length which is known as the neuterized map length. Now given a suitable map function $x=f(\theta)$, $\pi(\theta)$ could be derived as follows:

$$\pi(\theta) = \pi(f(\theta))\left|\frac{df(\theta)}{d\theta}\right| \qquad (5.7)$$

Similarly notice that, if the parameter of interest was x instead of θ then any of the above two priors could be used in term of x. A plot of the Haldane & Smith prior and the Renwick prior in term of both θ and x when the map function Eq(3) is used is provided in figure(5.1)(a,b) and (5.2)(a,b) respectively.

(5.3) Assessment and notation

The assessment of any method, as in the previous chapter, will be based on calculating the exact confidence coefficient, CC, and expected length, EL.

From table(1.7) and statement(1.5), the density function $p(r|\theta)$ for a data point r, if the mating type was either a phase known

FIGURE (5.1) Prior distributions of θ

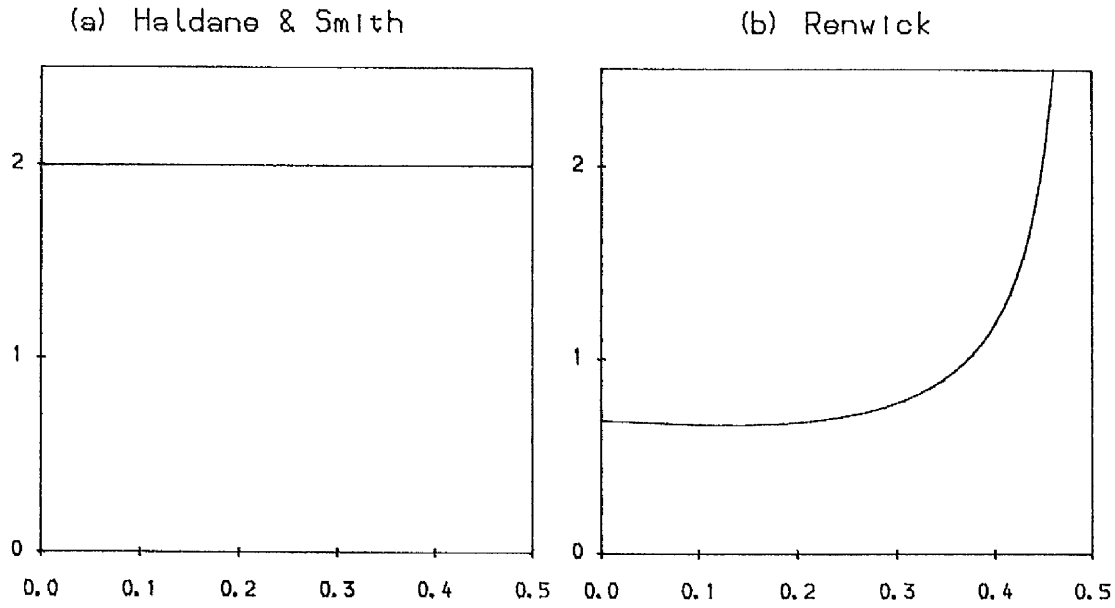
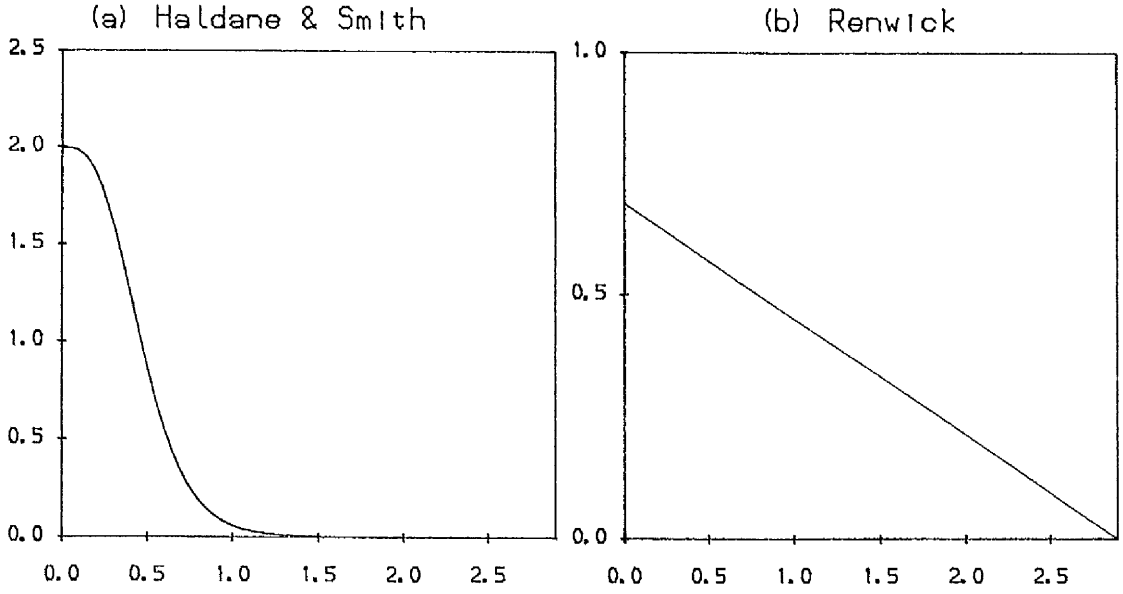


FIGURE (5.2) Prior distributions of x



or unknown double backcross respectively, would be the usual binomial distribution but with θ as the probability of success under the phase known situation and a function of θ, ϕ, under the phase unknown situation. For either density the sample space R would be $R=\{0,1,2,\ldots,n\}$, where n is the total number of offsprings.

Now if we let IE(l,r) be a 95% IE, produced by method l and data r, where:

l=1, when the likelihood approach as described in chapter 4, statement(4.1) and when the large sample approximation of (4.2) is used; (i.e) with p=1, $h'=-(0.5)x^2(1,0.95)=-1.92$.

l=2, when constructing a CCI Bayesian interval for θ with Haldane and Smith prior.

l=3, a HPD Bayesian interval with Haldane and Smith prior.

l=4, a CCI Bayesian interval with Renwick prior.

l=5, a HPD Bayesian interval with Renwick prior.

Also let:

$$T(l,r) = \begin{cases} 1 & \text{if } \theta_t \in IE(l,r) \\ 0 & \text{otherwise} \end{cases}$$

$$LL(l,r) = \theta_2(l,r) - \theta_1(l,r)$$

where $\theta_1(l,r)$, $\theta_2(l,r)$ are the lower and upper limit respectively of IE(l,r), then:

$$CC(l) = \sum_{r\in R} T(l,r)p(r|\theta) \qquad (5.9)$$

$$EL(l) = \sum_{r\in R} LL(l,r)p(r|\theta) \qquad (5.10)$$

The method(s) with the $CC\geqslant 0.95$ will be given more credit, and among them the one with the shortest EL will be preferred. Note that by using this method of assessment, we will be providing a frequentist assessment for a Bayesian interval which has been

derived from a completely different philosophy.

(5.4) Application

As in the previous chapter and for similar reason, we are going to apply the above methods when n=25, the map function, $f(\theta)$,is Eq(3) and when θ_t varies from 0.01 to 0.25 in step of 0.01, thus covering the area of practical interest.

(5.4.1) Phase known situation

(i) l=1

$$IE(1,r) = \{\theta \text{ such that } l(\theta)-l(\hat{\theta})>-1.92\}$$

where $l(\theta) = \text{const} + r\ln\theta + (n-r)\ln(1-\theta)$ for $0<\theta<0.5$

$$\hat{\theta} = \begin{cases} \frac{r}{n} & 0\leqslant \frac{r}{n} \leqslant 0.5 \\ 0.5 & \frac{r}{n} \geqslant 0.5 \end{cases}$$

$\theta_1(1,r)$ and $\theta_2(1,r)$, θ_1 and θ_2 for convenience, are just the roots of the following function, if within the feasable region [0.0,0.5],

$$H(\theta) = l(\theta) - l(\hat{\theta}) + 1.92$$

Numerical methods supplied by the Nag library routines have to be used to find θ_1 and θ_2.

(ii) l=2

$$IE(2,r) = \{\theta \text{ such that } \theta\in[\theta_1,\theta_2]\}$$

where θ_1 and θ_2 are as defined in (5.4) and

$$\pi(\theta|r) = \text{const } \theta^r(1-\theta)^{n-r} \qquad 0<\theta<0.5$$

By choosing the suitable constant $\pi(\theta|r)$ will be a truncated Beta distribution. Therefore finding θ_1 and θ_2 has been done by using the inverse of the incomplete beta function supplied by the Nag library routines.

(iii) l=3 (or 5)

$$IE(l,r) = \{\theta \text{ such that } \pi(\theta|r)>c\}$$

where c is chosen to satisfy (5.3). Therefore θ_1 and θ_2 will be the roots of the following function, $G(\theta)$, if within the feasable

region:

$$G(\theta) = \pi(\theta|r) - c$$

c, θ_1 and θ_2 have to be found numerically. Figure(5.3) is a flow chart showing the essential steps involved, given that $\pi(\theta|r)$ is unimodal. The following is a brief description of it:

(1)Find the mode of $\pi(\theta|r)$, θ_M, either analytically or numerically if not possible.

(2)If θ_M= 0.0 (or 0.5) the problem has to be redefined in the following way, θ_1=0.0 whereas θ_2 would be found such that

$$\int_0^{\theta_2} \pi(\theta|r)d\theta = 0.95$$

(3)θ_1 is in general between two carefully chosen points $Z^{(i)}$ and $Y^{(i)}$. Initially $Z^{(o)}$=0.0 and $Y^{(o)}$=θ_M.

(4)The average of $Z^{(i)}$ and $Y^{(i)}$ will be used as our iterative procedure which aims at finding θ_1. This procedure will certainly converge to the solution as long as the posterior density is unimodal,(i.e) $\theta^{(i)}$=$(Z^{(i)}+Y^{(i)})/2$. Also let f^i=$\pi(\theta_1^{(i)}|r)$.

(5)Now find $\theta_2{}^{(i)}$ which is the root of the following function

$$G(\theta) = \pi(\theta|r) - f^i \qquad \text{where} \quad \theta_M<\theta_2{}^{(i)}<0.5$$

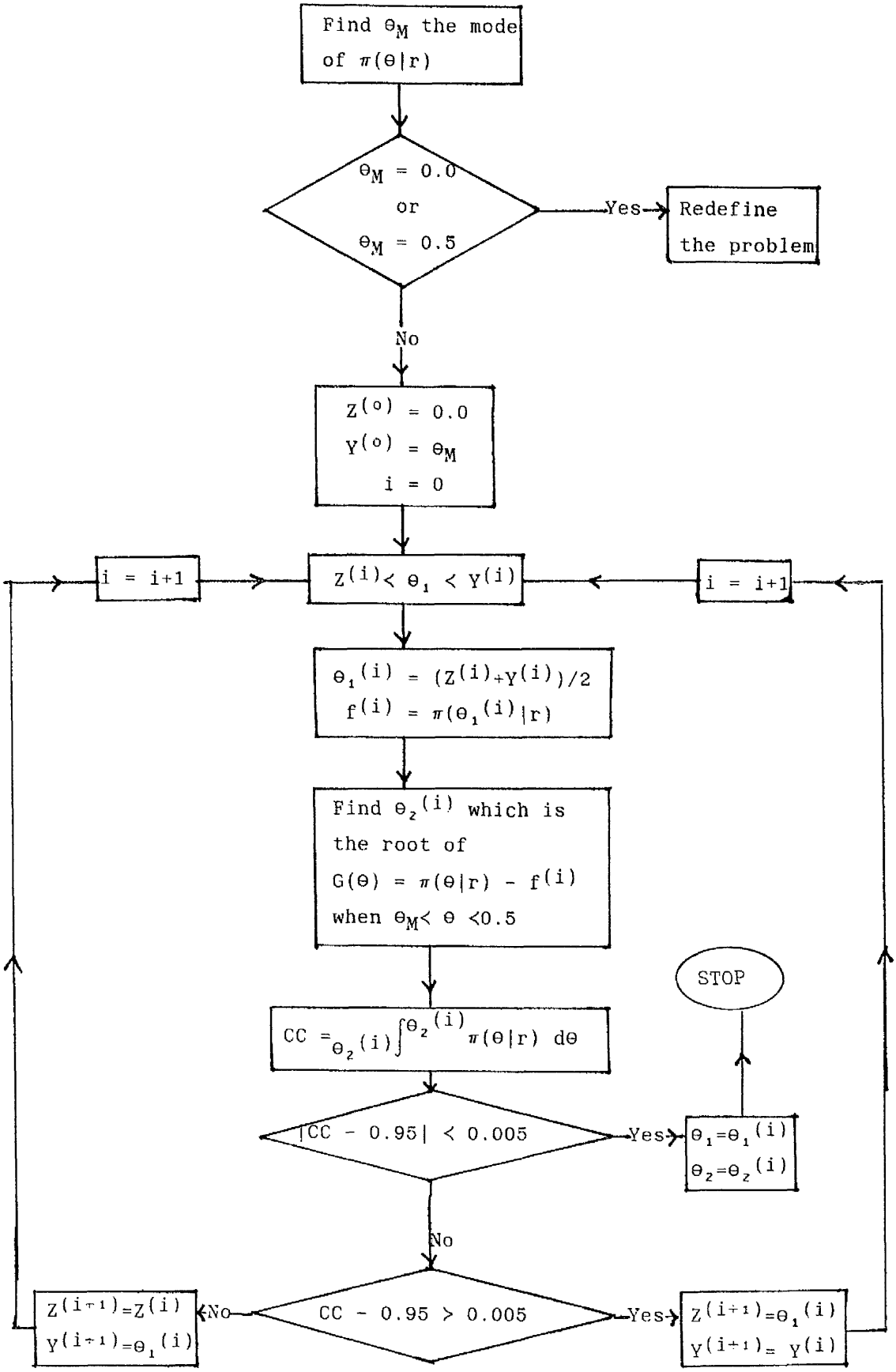
(6)Find CC, where

$$CC = \int_{\theta_1{}^{(i)}}^{\theta_2{}^{(i)}} \pi(\theta|r)d\theta$$

(7)If the accuracy required in finding CC is met, then stop, otherwise move to step (3) but with different values for $Z^{(i)}$ and $Y^{(i)}$ as described in the figure.

From graphical inspection of $\pi(\theta|r)$ when l=5, the posterior density has been found to be bimodal for many values of r. A possible half-way solution to this problem would be to find a one to one transformation, y=y(θ), such that the posterior density of the transformed variable y is unimodal. Note that, an exact 95%

Figure(5.3) Flow chart for finding the upper and lower limit of a 95% HPD Bayesian interval



HPD for y will lead to an exact 95% IE for θ although not necessarily a HPD one. This is so, because although any $100(1-\alpha)\%$ IE for θ as defined in (5.2,a) is invariant under any one to one transformation, the second requirement of a HPD interval for θ as defined in (5.2,b) will only be invariant if the transformation is linear. Both requirements of linearity and a resulting unimodal posterior density are contradictory. Fortunately, the requirement of unimodality could be achieved by using the map distance $x=f(\theta)$, which is a parameter of interest as much as θ is. Therefore a slight change in our previous plan which will produce a HPD for x instead of θ when the Renwick prior is used will also be of interest, (i.e) let:

l=5, be the method leading to a 95% HPD Bayesian interval for x when the Renwick prior is used. Also for consistency let:

l=4, be the method leading to a 95% CCI Bayesian interval for x when the Renwick prior is used. And for the sake of the comparison let:

l=6, be the method leading to a 95% IE for x when the likelihood approach is used. Actually the interval IE(6,r) will be exactly equal to the induced interval $f^{-1}(IE(1,r))$.

(iv) l=4

$$IE(4,r) = \{x \text{ such that } x \in [x_1, x_2]\}$$

where x_1 and x_2 are defined analogously to θ_1 and θ_2 in (5.4) and

$$\pi(x|r) = \text{const } (3-x)[f^{-1}(x)]^r[1-f^{-1}(x)]^{n-r}$$

where $0<x<\infty$ and $f^{-1}(x)$ has to be found numerically as described in the previous chapter. Again numerical method had to be adopted in order to find x_1 and x_2. The method used in this section is very similar to the prescribed one in section (iii).

(5.4.2) Phase unknown situation

Our aim is again to compare the IE for θ as produced by methods

l=1,2,3 and the IE for x as produced by methods l=4,5,6. The same steps of the technical calculation as in the phase known situation is going to be followed for this case, the following are only some points of difference:

(i)<u>l=1 (or 6)</u>

As $l(\theta)$ will be describing the usual binomial distribution but with probability of success equal ϕ, where $\phi=2\theta(1-\theta)$, then:

$$l(\theta) = \text{const} + r\ln(2\theta(1-\theta)) + (n-r)\ln(\theta^2+(1-\theta)^2)$$

So that by applying the usual analytical techniques for finding the MLE of the above ln likelihood, we found that (see appendix(5.1)) within the feasible region:

$$\hat{\theta} = \begin{cases} \frac{1}{2} & \text{if} \quad r > \frac{n}{2} \\ \frac{1}{2} - \frac{1}{2n}(n(n-2r))^{0.5} & \text{otherwise} \end{cases}$$

Figure(5.4)(a,b) is a plot of the ln likelihood under both of the above situations.

(ii)<u>l=2</u>

The posterior density $\pi(\theta|r)$ will obviously not be the incomplete beta distribution, so that finding θ_1 and θ_2 will be done by using our own numerical techniques which is very similar to the one described in figure(5.3) and which essentially depends on the unimodality of the density.

(iii)<u>l=3,4,5</u>

There is no technical difference between the phase known situation and this case for that matter. We are just providing the reader with a plot of the posterior density given Renwick prior, $\pi(x|r)$, for two different data points r, in figure(5.5)(a,b) which again shows the unimodality of these functions.

PHASE UNKNOWN SITUATION

FIGURE (5.4) Examples of the ln likelihood function

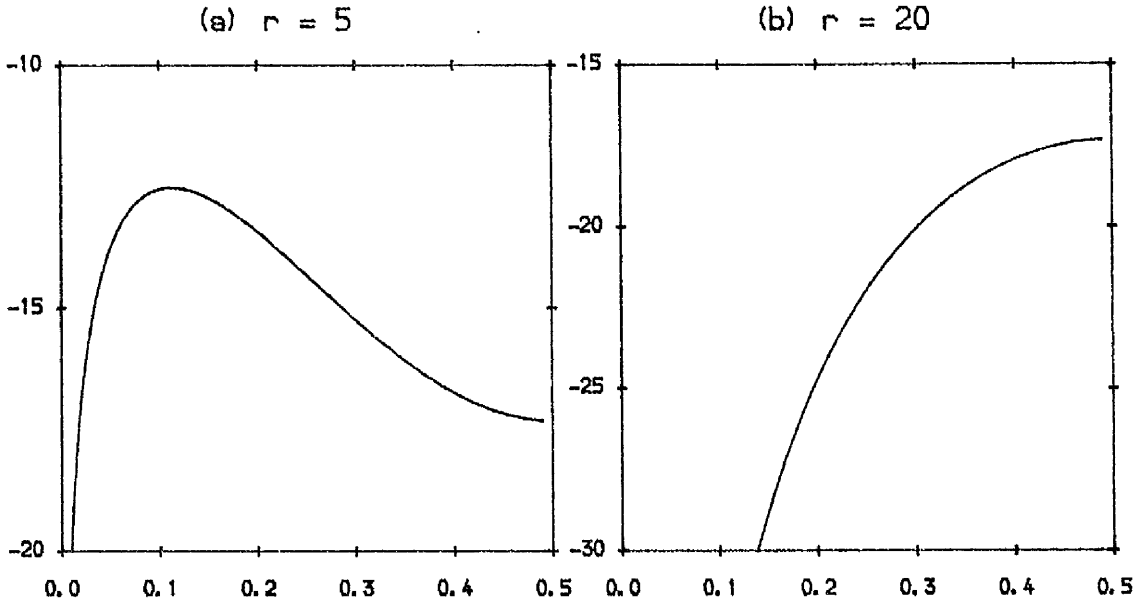
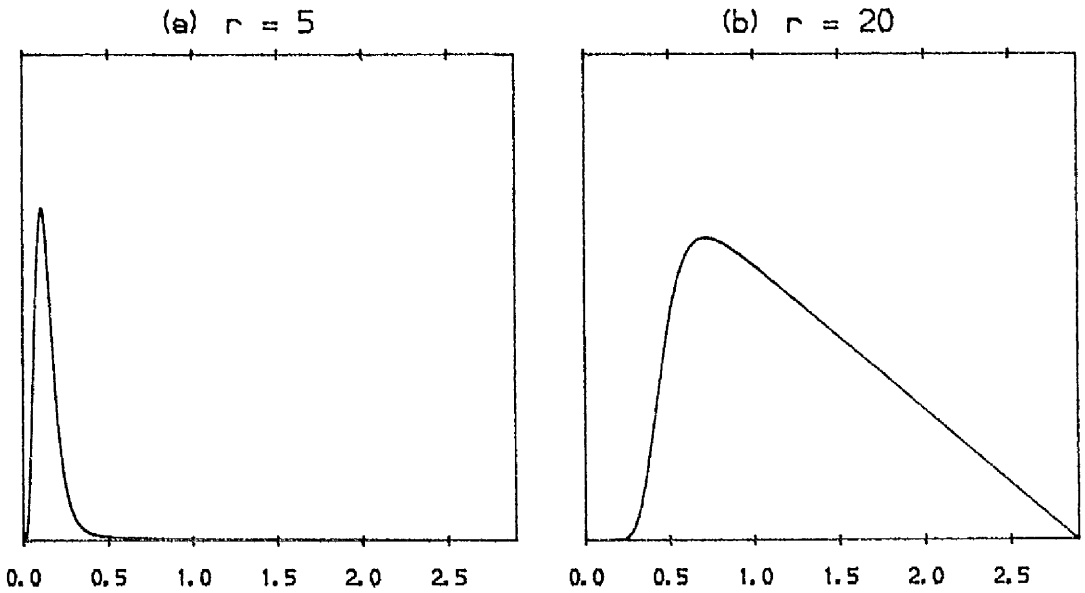


FIGURE (5.5) Examples of the posterior density of x when the second prior is used



(5.5) Results

The results of the above methods are shown in four different plots. A plot of the CC and the EL against the true parameter of interest θ for the three methods l=1,2,3 and two similar plots but against the parameter of interest x for the remaining three methods l=4,5,6. Figure(5.6) and (5.7) (a,b,c,d) shows the above four plots for the phase known and unknown situation respectively. In all of these plots, the black curve, the red curve and the green one represent the likelihood the CCI and the HPD intervals respectively.

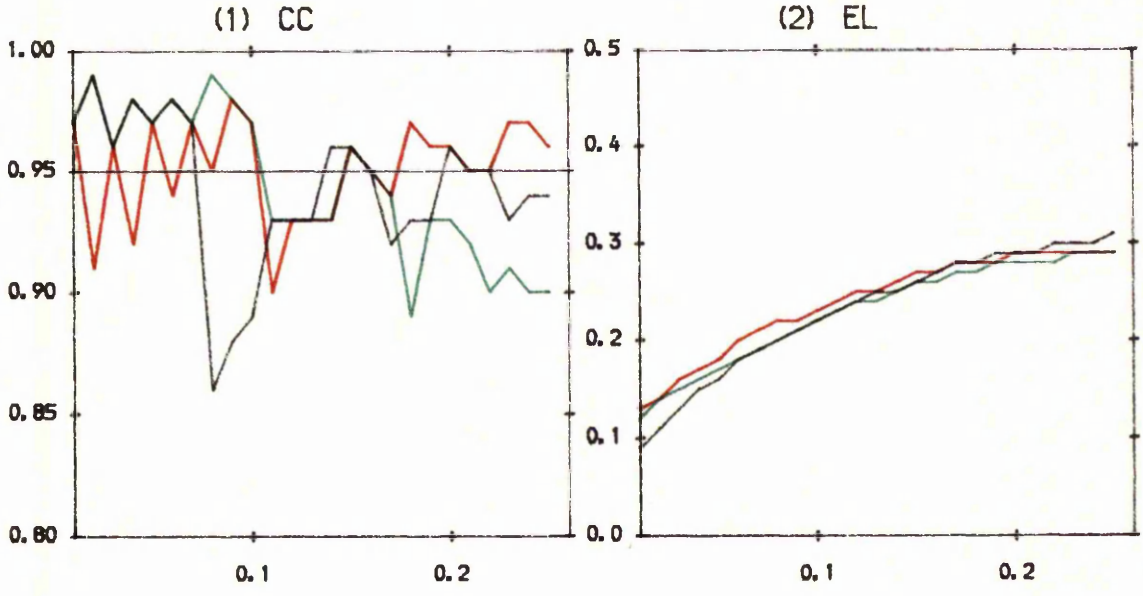
For the phase known situation and as far as the parameter θ is concerned, no method can be seen as a clear winner. All CC varies above and below the 0.95 threshold. Perhaps the CCI (with the first prior) could be seen as the most stable method with a minimum CC=0.90 and a maximum CC=0.98 and probably an average CC=0.95 for the studied range of θ_t, but as would be expected, it has the highest EL for almost all of this range. On the other hand both the HPD (with the first prior) and the likelihood methods are not as stable with a minimum CC=0.89 & CC=0.86 and a maximum CC=0.99 & CC=0.99 respectively, also notice that the difference between their EL is almost negligible for the studied range of θ_t. As for the parameter x and as far as the CC is concerned the HPD seems to be the best method for almost all the studied range of x_t its minimum CC=0.94, maximum CC=1.0 and average CC=0.97 for the corresponding range of x_t. The likelihood method on the other hand has clearly the least expected length.

For the phase unknown situation very similar comments could be made, again the CCI interval when the parameter of interest is θ seems to be the most stable method as far as the CC is concerned but which also gives the highest EL for this case. On the other

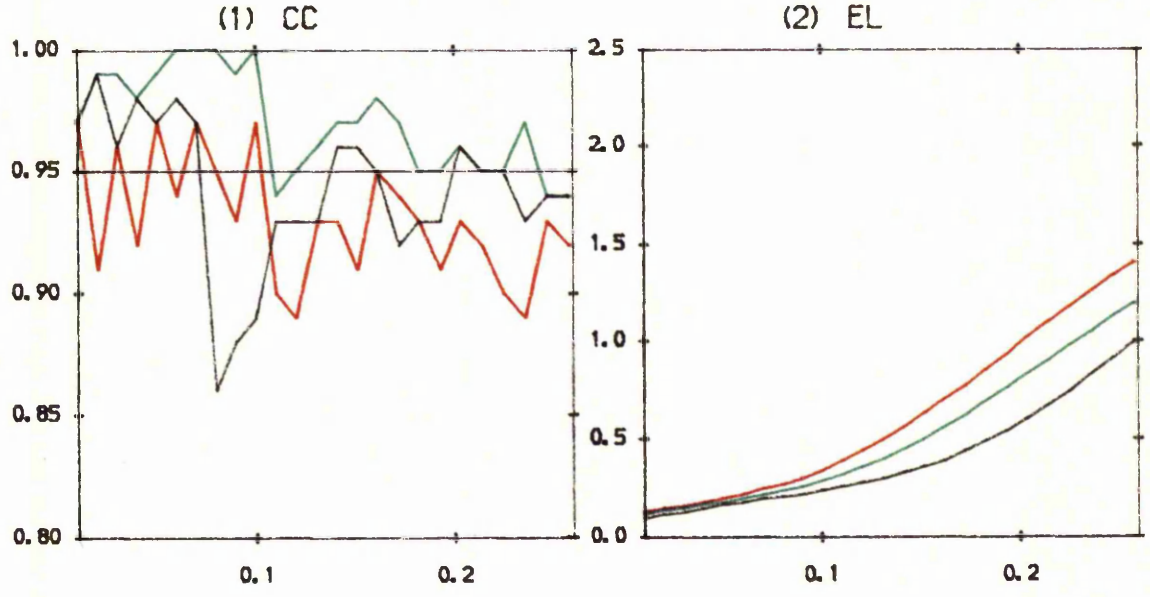
PHASE KNOWN SITUATION

FIGURE (5.6) A comparison between the likelihood & Bayesian approaches in construction IE

(a) θ is the parameter of interest & first prior is used



(b) x is the parameter of interest & second prior is used

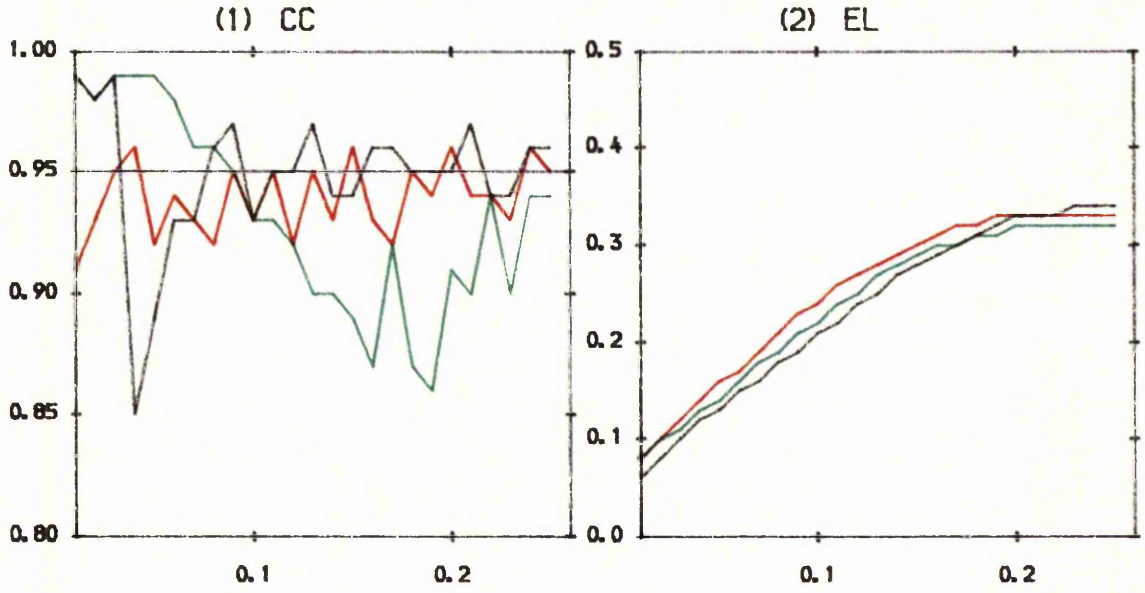


Likelihood
Bayesian CCI
Bayesian HPD

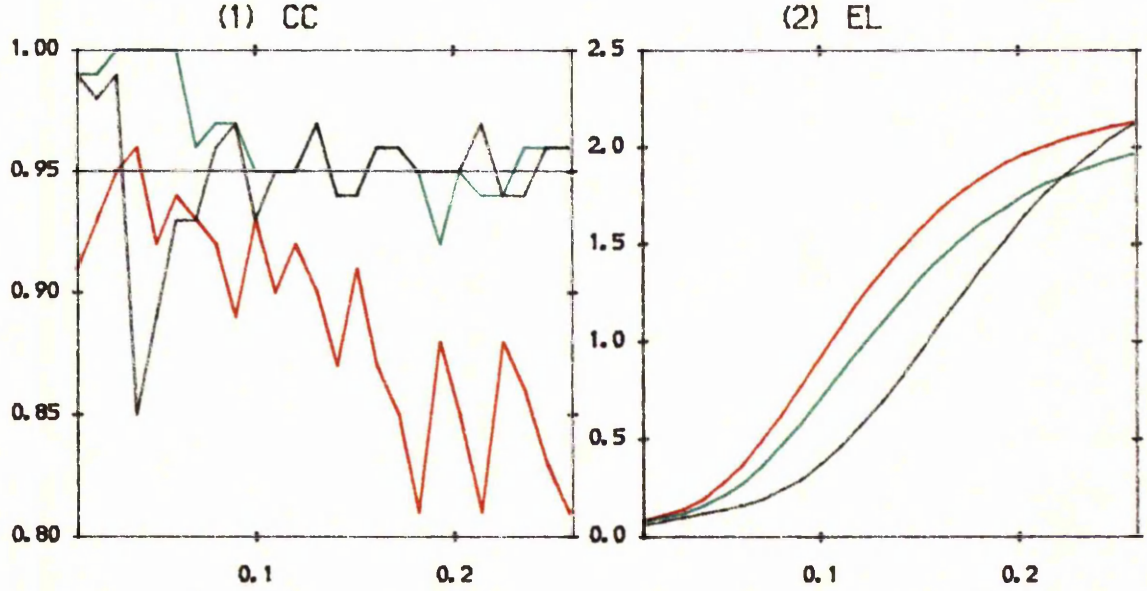
PHASE UNKNOWN SITUATION

FIGURE (5.7) A comparison between the likelihood & Bayesian approaches in construction IE

(a) θ is the parameter of interest & first prior is used



(b) x is the parameter of interest & second prior is used



hand this same method is clearly the least favourable method if the parameter of interest was x, notice also that for this situation the HPD seems again to be the best method as far as the CC is concerned whereas the likelihood method has clearly the least EL for almost all the studied range of x_t.

Actually the reasonable behaviour of the HPD interval for the parameter x could be understood by looking back at the plot of the second prior distribution, figure(5.2)(b), which is clearly favouring smaller values of x and which therefore probably will lead to a HPD Bayesian intervals with a high CC for small values of x_t as well as a large EL for large values of x_t. This same prior could partially explain the drastic behaviour of the CCI Bayesian intervals for the phase unknown situation and the not so drastic one for the phase known situation, but still the least favourable behaviour as far as both CC and EL are concerned. Actually the form of the prior as well as the likelihood function would lead to a highly skewed posterior density, especially for the phase unknown situation, making a central Bayesian interval not a very sensible choice; and as the skewness is very much toward the large values of x_t ((eg) $x_t>0.7$ for figure(5.5)(b)) this could probably lead to a wide interval capturing most of the large values of x_t and not so much of the smaller values. This last comment would suggest that the red curve in both figure(5.6)(b,1) and (5.7)(b,1) would be much higher for large x_t (this is actually true, plot not shown here). Also notice that although the difference between the two Bayesian intervals seems quite large for any of the two parameters of interest and as far as the studied range is concerned, it is a theorem that for the whole range of θ or x any Bayesian interval will have an average CC equal 0.95 exactly, (i.e) if CC is seen as a random variable

of θ for example then:

$$E_\theta(CC(\theta)) = \int_0^{0.5} CC(\theta)\pi(\theta)d\theta = 0.95$$

The proof of this theorem is quite an easy and short one which is given here for θ as follows:

From (5.9) $CC(\theta)$ could be rewritten as follows:

$$CC(\theta) = \sum_{r \in R} T(\theta,r)\ p(r|\theta) \qquad (5.9)(a)$$

where using a Bayesian philosophy the indicator variable $T(\theta,r)$ in (5.9) would be seen as a random variable, as it depends on the random variable θ, and which also depends on the given data point r, hence the notation $T(\theta,r)$. So that from (5.9)(a), $CC(\theta)$ could be seen as the conditional expectation of $T(\theta,r)$ given θ, (i.e):

$$CC(\theta) = E_r(T(\theta,r)|\theta)$$

So that:

$$E_\theta(CC(\theta)) = E_\theta(E_r(T(\theta,r)|\theta))$$
$$= E_r(E_\theta(T(\theta,r)|r))$$

But for any given r:

$$E_\theta(T(\theta,r)|r) = \int_0^{0.5} T(\theta,r)\pi(\theta|r)d\theta$$
$$= \int_{\theta_1}^{\theta_2} \pi(\theta|r)d\theta \qquad (5.11)$$

But from the definition of any Bayesian interval, equation (5.11) should be equal to 0.95 so that:

$$E_\theta(CC(\theta)) = E_r(E_\theta(T(\theta,r)|r)) = E_r(0.95) = 0.95.$$

(5.6) Three loci situation & the Bayesian approach

(5.6.1) Introduction

A more interesting comparison between the Bayesian and likelihood approach could be achieved by applying both approaches to the more complicated case of the three loci situation. As seen in the previous chapter, by choosing a certain map function and given a certain mating type, we will be tackling a two

dimensional problem with the unknown parameter vector $\underline{\theta}=(\theta_1\ \theta_2)^T$. The likelihood interval for $\underline{\theta}$ will be completely specified through the density function of the data vector $\underline{r}$, the HPD Bayesian interval, as defined in (5.3), on the other hand, will still need a sensible choice of the prior distribution of $\underline{\theta}$.

(5.6.2) Prior distribution of θ

A generalisation of the previously introduced prior distributions are presented in this section. A generalisation of the Haldane and Smith prior will be based on the assumption of two independent uniform distributons for each of the recombination fraction θ_1 and θ_2, (i.e)

$$\pi(\underline{\theta}) = 4 \qquad \text{for} \quad 0<\theta_1<0.5,\ 0<\theta_2<0.5$$

Renwick et al(1971) introduced a generalisation to his previous prior to suit the situation of three autosomal loci, the generalised prior was based on similar assumptions to the one introduced in section(5.2.3). Again, the key assumption (b) has to be adjusted here to suit our case of three loci known to be on the same chromosome. Therefore (b) will be as follows; assume that any of the three loci A, B or C are equally likely to occur at any point on the whole chromosome of length L, (i.e) the distribution of any of their position is uniform[0,L]:

$$\pi(A,B,C) = (1/L^3) \qquad \text{for} \quad 0<A<L,\ 0<B<L,\ 0<C<L$$

Now let:

$x_{ab} = B-A$

$x_{bc} = C-B$

$u = A$

By using the jacobian transformation and then integrating over the whole range of u we can arrive at the prior distribution of x_{ab} and x_{bc}. Firstly:

$$\pi(x_{ab},x_{bc},u) = (1/L^3) \qquad \text{where} \quad 0<u<L$$

$$0<x_{ab}+u<L$$

$$0<x_{ab}+x_{bc}+u<L$$

$$\text{and} \quad -L<x_{ab}<L$$

$$-L<x_{bc}<L$$

$$-L<x_{ab}+x_{bc}<L$$

Secondly by integrating carefully over the whole range of u, the prior distribution of x_{ab} and x_{bc} will have six different possibilities corresponding to six different possible orders:

$$\pi(x_{ab},x_{bc}) = \begin{cases} (L-x_{ab}-x_{bc})/L^3 & \text{where } 0<x_{ab}<L \text{ ; } 0<x_{bc}<L-x_{ab} \\ & \text{corresponding to order ABC} \\ (L-x_{ab})/L^3 & \text{where } 0<x_{ab}<L \text{ ; } -x_{ab}<x_{bc}<0 \\ & \text{corresponding to order ACB} \\ (L+x_{bc})/L^3 & \text{where } 0<x_{ab}<L \text{ ; } -L<x_{bc}<-x_{ab} \\ & \text{corresponding to order CAB} \\ (L+x_{ab}+x_{bc})/L^3 & \text{where } -L<x_{ab}<0 \text{ ; } -L-x_{ab}<x_{bc}<0 \\ & \text{corresponding to order CBA} \\ (L+x_{ab})/L^3 & \text{where } -L<x_{ab}<0 \text{ ; } 0<x_{bc}<-x_{ab} \\ & \text{corresponding to order BCA} \\ (L-x_{bc})/L^3 & \text{where } -L<x_{ab}<0 \text{ ; } -x_{ab}<x_{bc}<L \\ & \text{corresponding to order BAC} \end{cases}$$

As a check to the derivation of the above joint prior, we calculated the marginal prior of x_{ab} which will be as follows:

$$\pi(x_{ab}) = \begin{cases} (L-x_{ab})/L^2 & 0<x_{ab}<L \\ (L+x_{ab})/L^2 & -L<x_{ab}<0 \end{cases}$$

where the first arm, for example, of $\pi(x_{ab})$ is calculated by adding and integrating the first three arms of $\pi(x_{ab},x_{bc})$ over the corresponding range of x_{bc}. But because the order of the two loci A and B are irrelevant to the investigation concerning linkage then let $x_1 = |x_{ab}|$, so that

$$\pi(x_1) = 2(L-x_1)/L^2 \qquad 0<x_1<L$$

which is equivalent to the prior distribution of the map distance between two loci on the same chromosome as defined in (5.6) and derived in section(5.2.3).

Now let $x_1 = |x_{ab}|$ and $x_2 = |x_{bc}|$, where x_1 and x_2 are the map distance between the loci A & B and B & C respectively. Then:

$$\pi(x_1,x_2) = \begin{cases} 2(L-x_1-x_2)/L^3 & 0<x_1+x_2<L \\ 2(L-x_1)/L^3 & 0<x_1<L \\ 2(L-x_2)/L^3 & 0<x_2<L \end{cases}$$

Actually this prior distribution is the joint prior of the two map distances and a certain gene order. Therefore if we were to assume a known gene order, for example order ABC, then:

$$\pi(x_1,x_2|ABC) = 6(L-x_1-x_2)/L^3 \qquad 0<x_1+x_2<L$$

Finally using the appropriate map function, $\pi(\underline{\theta}\cap order)$ or $\pi(\underline{\theta}|order)$ could be easily derived using the above prior for $\pi(x_1,x_2\cap order)$ and $\pi(x_1,x_2|order)$ respectively.

To compare the Bayesian approach with the different likelihood intervals produced in the previous chapter, we have to calculate the exact CC for the joint and marginal Bayesian intervals as well as the EL for the marginal ones.

(5.6.3) Marginal HPD Bayesian interval

The marginal distribution of θ_1 (or θ_2) could be easily derived by integrating, usually numerically, $\pi(\underline{\theta}|\underline{r})$ over the whole range

of θ_2 (or θ_1). Therefore a 95% marginal HPD Bayesian interval for θ_1 would be defined as follows:

$$L\theta_1 \int^{U\theta_1} {}_0\int^{0.5} \pi(\underline{\theta}|\underline{r}) d\theta_2 \; d\theta_1 = 0.95$$

such that: $${}_0\int^{0.5} \pi(L\theta_1, \theta_2|\underline{r}) \; d\theta_2 = {}_0\int^{0.5} \pi(U\theta_1, \theta_2|\underline{r}) \; d\theta_2$$

Numerical methods, very similar to the previously introduced one in figure(5.3), will have to be used in order to find $L\theta_1$ and $U\theta_1$.

(5.6.4) Joint HPD Bayesian interval

From the definition of this interval in (5.3), we can see that, given that $\pi(\underline{\theta}|\underline{r})$ has been derived, the interval will then depend entirely on the value of the constant c. With multidimensional problems the evaluation of c is not an easy task. An approximate evaluation of c could be achieved by simulating (large) number I of points $\underline{\theta}^S$ arising from the posterior density $\pi(\underline{\theta}|\underline{r})$. For each of these simulated points we can calculate the random variable $\pi(\underline{\theta}^S|\underline{r})$. The distribution function of $\pi(\underline{\theta}^S|\underline{r})$, $F(\pi)$, would be then used in order to find the lower 5% quantile of $\pi(\underline{\theta}^S|\underline{r})$, thus providing an approximate evaluation of c.

CHAPTER SIX: Predective estimate of the probability of risk - an example

6.1 Introduction

An important application of linkage studies is the use of linkage information in genetic counselling. For some families which are affected by a certain hereditary disease, calculating the probability that an unborn child (fetus) is carrying the disease gene is of major importance.

Genetic linkage studies which have been carried out throughout history of the subject, have established linkage between many disease genes and marker genes. For linked diseases and given the family history and mode of inheritance at the disease and marker loci, the above probability will be function of the recombination fraction(s) $\underline{\theta}$. If we are considering a family of (m-1) members with x_i and g_i denoting the phenotype and genotype of the i^{th} individual respectively, then the probability of the m^{th} unborn individual being at risk is equal to the conditional probability of him having a genotype g_m, where g_m incorporates the disease, given all phenotipic information in the pedigree, (i.e) this probability will be equal to $P(g_m|x_1..x_m^-)$. Note that x_m^- denotes the incomplete phenotype of the m^{th} individual, inspected by analysing the amniotic fluid of the pregnant mother, White(1984), which will lead to phenotipic information only about his marker gene(s). Seen as a function of $\underline{\theta}$ this probability could be written as $R(\underline{\theta})$. If, from previous linkage studies the MLE of $\underline{\theta}$ has been established to be $\hat{\theta}$, then by using the likelihood approach a point estimate of $R(\underline{\theta})$ will be given by the MLE $R(\hat{\theta})$. The aim of this chapter was to introduce the Bayesian approach for providing a point estimate of this function. Later on we

found out that this approach was previously introduced by Renwick et al(1971). Neverthless a short discussion plus application, using an example from the genetic literature, of this approach is given here.

6.2 Bayesian approach and probability of risk

Aitchson and Dunsmore(1975) discussed in detail the nature of statistical prediction analysis and its various applications. In their description of the problem they stated that "an essential feature of statistical prediction analysis is that it involves two experiments e and f. From the information which we gain from the performance of e, the informative experiment, we wish to make some reasoned statement concerning the performance of f, the future experiment". Here both experiments e and f are linked through a common unknown parameter $\underline{\theta}$. In other words, the problem could be described as follows. If the future experiment f has outcome y with sample space Y and a class of possible density functions $\{p(y|\theta) : \theta\epsilon\Theta\}$ on Y, where the parameter space Θ is assumed known but the true parameter θ_t is unknown, then the nature of the predictive problem is the uncertainty about θ_t and the final objective is to assess the plausibility of the unknown outcome y through the study of the plausibility of θ.

Two steps would therefore be involved. The first step is to assess the plausibility of θ using two sources of information, a prior information expressed in term of a known prior density $\pi(\theta)$ on Θ and the informative experiment e. If X is the sample space of e than the class of density functions for e can be denoted by $\{p(x|\theta) : \theta\epsilon\Theta\}$ on X. Using Bayes theorem the posterior density function of θ is given by:

$$P(\theta|x) = \frac{\pi(\theta)p(x|\theta)}{p(x)} \qquad (6.1)$$

The second step involves the calculation of the predictive density function $p(y|x)$ for y given $\pi(\theta)$ and x which is equal to:

$$p(y|x)= \int_{\Theta} p(y|\theta)p(\theta|x)d\theta \qquad (6.2)$$

The above approach in determining a predictive distribution for Y could be adapted to provide a predictive estimate of the probability of risk. The counselling problem would therefore be described as follows. The previous family trees which provided information about θ through the different phenotypes of their members at the disease and marker loci could be seen as being the informative experiment e with sample space $\underline{Z}$ –all possible phenotypes–, parameter space Θ and a class of density functions $\{p(\underline{z}|\theta) : \theta \in \Theta\}$ on $\underline{Z}$. By choosing a suitable prior $\pi(\theta)$, the posterior density $\pi(\theta|\underline{z})$ could be calculated using formula(6.1). Another family tree is observed with outcome $\underline{x}=(x_1 x_2 .. x_{\bar{m}})$, sample space $\underline{X}$ and with the same parameter space Θ. Our main concern now is not the predictive density of a future outcome but rather the predictive probability of a certain event, merely that individual m carries the disease gene given his family tree $\underline{x}$, $\pi(\theta)$ and the previous data $\underline{z}$.

If we are dealing with a dominant hereditary disease with two possible alleles, D denoting the abnormal dominant allele and N the normal recessive one, then the event of interest R will be that the genotype of individual m at the disease locus is either D/D or D/N, (i.e) R={D/D,D/N}. As seen in section(6.1), the probability of this event given the family tree $\underline{x}$ will depend on the parameter θ. Therefore the predictive probability could be calculated in a manner analogous to (6.2) as follows:

$$P(R|\underline{x},\underline{z})= \int_{\Theta} P(R|\underline{x},\theta)p(\theta|\underline{z})d\theta \qquad (6.3)$$

Formula(6.3) is only true if $\underline{x}$ does not provide any information

about θ. If it does then the posterior density of θ should be calculated using the information provided from $\pi(\theta)$, $p(\underline{z}|\theta)$ and $p(\underline{x}|\theta)$ which will lead to $p(\theta|\underline{z},\underline{x})$. The predictive probability will then be:

$$P(R|\underline{x},\underline{z})= \int_{\theta}P(R|\underline{x},\theta)p(\theta|\underline{z},\underline{x})d\theta \qquad (6.4)$$

As $P(R|\underline{x},\theta)$ could be seen as a function of θ, $R(\theta)$,then

$$P(R|\underline{x},\underline{z})= \int_{\theta}R(\theta)p(\theta|\underline{z},\underline{x})d\theta$$

$$= E(R(\theta)|\underline{z},\underline{x}) \qquad (6.5)$$

6.3 Application

Morton(1956) studied linkage between the Rh blood group gene and Elliptocytosis, a dominant disease gene. The marker gene Rh is determined mainly by 3 codominant alleles R_1, R_2 and r. The disease gene is determined by the rare dominant disease allele El and the common normal recessive allele el. Fourteen independent family trees were studied, in his paper Morton called them pedigree 1,2,..,7, R, B, A.E, S.S, M.K, J.P.N and J.M.L respectively. He derived the likelihood function of each family using the technique described previously in section(1.5-A). Using all pedigrees and a generalised likelihood ratio test, linkage was clearly found to be significant ($(2\ln 10)Z(\hat{\theta})=34.2$). Nevertheless by separately analysing each family pedigree, some families showed clear linkage while others offered no evidence on linkage. Using what Morton called a heterogeneity test, which is just a generalised likelihood ratio test for testing

, $H0:\theta_1=\theta_2=...=\theta_{14}=\theta$ against

H1: H0 not true

where θ_i is the recombination fraction of the i^{th} family, H0 was clearly rejected. Most of the variation between families was found out to be mainly between two groups of pedigrees, group 1

which consists of pedigree 3, 4, 5 and R and group 2 which consists of pedigree 2, A.E and J.P.N. Also homogeneity within each group was not significant. One explanation of the above result, which was given by Morton(1956), was that elliptocytosis depends on two loci, the first of which is closely linked to the Rh locus (as in pedigree 3, 4, 5 & R) while the second is in a different linkage group (as in pedigree 2, AE & J.P.N).

Given that the above explanation is true, let us assume that one of the two simple families which are shown in figure(6.1) came for counselling about the risk that their unborn child is carrying the disease gene El. To answer their inquiry we can either adopt a likelihood or a predictive approach, but in either case we will be faced with the extra problem of heterogeneity. We can either assume that the new family belongs to the linked group of families and analyse the data accordingly or more appropriately include the uncertainty about the type of family (being in the linked group or not) when analysing the data. The former procedure will probably lead to a biased estimate of the probability of risk. But as our main interest is to compare the likelihood and the predictive approach we are going to use both methods.

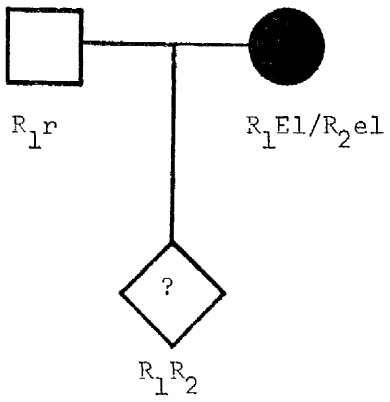
First method:

Let $\underline{x}_1, \underline{x}_2$ be the phenotypes outcome of the 1st and 2nd family tree described in figure(6.1) respectively. The 1st family is a phase known mating of the form $R_1El/R_2el \times R_1el/rel$. Their unborn child must have received the R_1 gene from the unaffected father and the R_2 gene from his affected mother. The probability that he also received the disease allele El from his mother is equal to the recombination fraction θ, (i.e)

$$P(R|\underline{x}_1,\theta) = \theta.$$

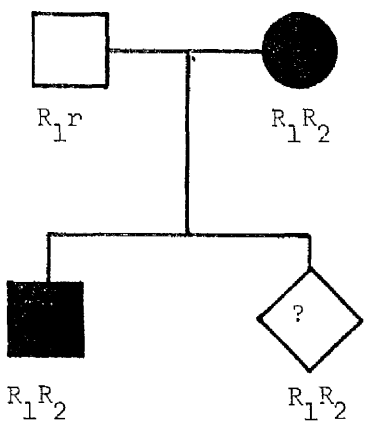
FIGURE(6.1)

First family

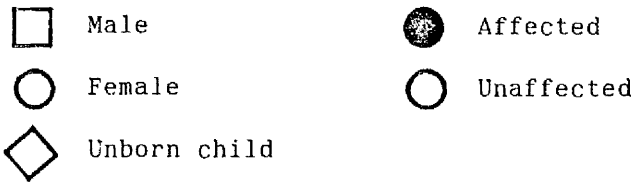


(The affected mother phase is known).

Second family



(The affected mother phase is unknown, either R_1El/R_2el or R_1el/R_2El)



As for the 2nd family the phase of the affected mother is unknown we have one of the possible two equally likely mating $R_1El/R_2el \times R_1el/rel$ or $R_1el/R_2El \times R_1el/rel$. Taking both matings into account $P(R|\underline{x}_2,\theta)$ could easily be seen to be equal to:

$$P(R|\underline{x}_2,\theta)= \theta^2+(1-\theta)^2 .$$

To apply the predictive approach we chose to use Haldane and Smith's prior (i.e)

$$\pi(\theta)=2 \qquad 0<\theta<0.5$$

Also note that both family trees $\underline{x}_1,\underline{x}_2$ do not add any further information about the plausibility of θ and therefore will not be used in deriving the posterior density of θ.

Second method:

Smith(1963) provided a model which account for the heterogeneity problem. He assumed that a proportion λ of families show linkage between the two gene loci under question while in the remaining proportion, $(1-\lambda)$ of families, the two loci are unlinked. Therefore the likelihood of a certain family i with phenotype outcome $\underline{x}_i$ could be expressed in terms of θ and λ as follows:

$$L_i(\theta,\lambda)= \lambda P(\underline{x}_i|\theta)+(1-\lambda)P(\underline{x}_i|0.5)$$

Under this model, the risk probability for the fetus in the 1st and 2nd family will be equal to:

$$P(R|\underline{x}_1,\theta)= \lambda\theta+0.5(1-\lambda)$$

$$P(R|\underline{x}_2,\theta)= \lambda(\theta^2+(1-\theta)^2)+0.5(1-\lambda)$$

To apply the predictive approach under this model we may assume that all values of θ and λ within their range are equally likely, as has also been suggested by Smith, therefore

$$\pi(\theta,\lambda)= 2 \qquad 0<\theta<0.5$$

$$0<\lambda<1$$

6.4 Results

Aitchison and Dunsmore(1975) pointed out that the likelihood and the predictive approach will be in good agreement if $p(\theta|\underline{z})$ is highly concentrated on $\hat{\theta}$. By the usual large sample arguments this will be the case when there is a substantial past experience. Therefore we suspect that by using all the available family pedigrees, the predictive and likelihood approach will give close results. This led us to present the analysis of the data as if being done sequentially or step by step as follows. For the first method, we have all in all four previous family pedigrees, pedigree 3, 4, 5 and R. Table(6.1)(a) presents the likelihood and predictive probability of risk for our two family trees when the previous data used, first of all, is only pedigree 3 and then pedigree 3 and 4 and so on until we use all previous data. As suspected the disagreement at the beginning of the table is much higher than at its end. Also the predictive approach shows a more stable estimate. Actually the least estimated value of the risk probability using the likelihood approach and given the 1^{st} family tree is 0.0 whereas the largest is 0.039. By using the predictive method these values are 0.046 and 0.059 respectively. This stability is probably due to the fact that at any stage of the analysis, the predictive approach weights the possible probabilities $P(R|\underline{x}_1,\theta)$ according to the plausibilities of the various θ in contrast to the likelihood approach which takes no acount of the sampling variability of the estimator $\hat{\theta}(\underline{z})$. Similar comments could be said about the result of the 2^{nd} method. These results are shown in table(6.3)(b), in which we have fourteen steps as we have all fourteen pedigrees to use as previous data. Also notice that, at step 1 in the table, the MLE of θ and λ were equal 0.00 and 1.0 respectively and had changed

Table(6.1) The likelihood and predictive probability of risk

(a)When the families are assumed to be in the linked group

Previous data	LIKELIHOOD		PREDICTIVE	
Pedigree	$\hat{\theta}=P(R\|x,\hat{\theta})$	$P(R\|x_2,\hat{\theta})$	$P(R\|x_1,z)$	$P(R\|x_2,z)$
3	0.000	1.000	0.048	0.913
3+4	0.031	0.940	0.059	0.892
3+4+5	0.039	0.925	0.056	0.896
3+4+5+R	0.032	0.939	0.046	0.914

(b)When the families' group are not known

Previous data	LIKELIHOOD		PREDICTIVE	
Pedigree	$P(R\|x_1,\hat{\theta},\hat{\lambda})$	$P(R\|x_2,\hat{\theta},\hat{\lambda})$	$P(R\|x_1,z)$	$P(R\|x_2,z)$
1	0.000	1.000	0.357	0.600
1+2	0.371	0.533	0.403	0.557
1+2+3	0.211	0.789	0.268	0.703
1+2+3+4	0.169	0.811	0.226	0.742
1+...+5	0.143	0.830	0.196	0.770
1+...+6	0.134	0.839	0.190	0.755
1+...+7	0.116	0.857	0.171	0.797
1+...+R	0.099	0.879	0.149	0.822
1+...+B	0.101	0.878	0.152	0.820
1+...+A.T	0.156	0.825	0.190	0.786
1+...+S.S	0.188	0.792	0.229	0.780
1+...+M.K	0.194	0.785	0.216	0.757
1+...+J.P.N	0.230	0.752	0.245	0.733
ALL	0.248	0.734	0.256	0.720

to become 0.036 and 0.542, at the final step of the analysis.

CHAPTER SEVEN: Discussion

Once linkage has been established between a group of n loci, where $n \geq 3$, then another major problem would confront the investigator(s), that of establishing the order of this group of loci. With n=3, only three orders can be tested against each other; with n>3, the problem becomes extensively complicated as the number of the tested orders, (n!/2), increases tremendously.

In chapter two and three of this study, we have been mainly interested in studying the statistical performance of a certain criterion, mentioned by Ott(1985) to distinguish between two gene orders. This criterion has been applied to test the three different orders obtained when dealing with a three loci situation. But because of the special set up of the problem, for which we are testing three hypotheses simultaneously and where all of which are of the same dimensionality, the usual assessment which is based on the concept of a significance level and power, was in our view not suited. Actually our assessment of this criterion, was based on the probability of it leading to a right, a wrong or to an inconclusive decision, denoted by P_1, P_W and P_4 respectively.

By using simulated data from phase known triple backcross families with three codominant alleles, and when the Haldane map function was assumed, we found out that the interesting range of N, the sample size, to be studied was [15,30]. A lower value of N will lead to a high probability of inconclusive decision, whereas a higher value will lead to a high probability of a right one. Using this result, we started out a more elaborate simulation , for N=20, upon which we estimated the three interesting probabilities, P_1, P_W and P_4,as a function of $\underline{\theta}=(\theta_1\ \theta_2)^T$ when the

true map function was Eq(3) (Morton(1987)) and when the assumed one was either the Haldane, Kosambi or Eq(3). We noticed that, for any assumed map function, $\hat{P}_4$ was very large when either of the θ_i was near 0.5 and then began to decrease as either of them decreases to reach a minimum value roughly at $\theta_1=\theta_2=0.12$, this comment would exactly be reversed to suit $\hat{P}_1$. On the other hand the smallness of $\hat{P}_W$ for all of the range of $\underline{\theta}$, was the noticeable feature of this estimated probability; the maximum estimated value was roughly equal 0.0055, 0.012 and 0.023 under the three assumed map functions respectively. This result seems to suggest that using the prescribed criterion, section(2.2) page 54, to choose one of the tested orders is quite a safe one to use. Also, as far as the assumed map functions are concerned, we noticed that both $\hat{P}_1$ and $\hat{P}_W$ increased with the correctness of this function. Actually at the maximum value of $\hat{P}_1$, the ratio between the Haldane and Eq(3) was about 2:3.

Morton et al(1986) used all $\binom{9}{3}$ trios of loci from a nine point data on the X chromosome of the *D.melanogaster* published by Morgan et al (1935), to calculate the ratio of the support, S, for the maximum order provided by using the Haldane map function against a more realistic one, the Rao et al(1979) with p=0.35; where the support, S, was defined as the maximised ln likelihood ratio between the maximum order and the next maximum one. The value of this ratio was found to be equal 0.66. This loss of support for the correct order, which seems to correspond to the maximum one in that example, led them to believe that a more realistic map function than the Haldane would be then required for testing orders. Lathrop et al(1985) calculated the relative odds of the maximised likelihood functions between the three

different orders for a certain three point human data, when either the Haldane, the Kosambi or even complete interference was assumed. From that example and others, they noticed that by assuming an interference level other than the Haldane, it had the effect of increasing the relative odds in favour of the maximum order, consequently they concluded that the assumption of no interference was a conservative one. They also commented in a later paper(1987), that larger odds should not be interpreted as an increased evidence for the maximum likelihood order without calculation of significance levels.

Actually most of the above comments, seem to agree with our result. The conservatism of the Haldane map function had certainly been revealed in our result as having a smaller $\hat{P}_W$ compared to the other assumed functions. Also, it seems that the loss of support for the correct order, measured by Morton, had revealed itself in a 2:3 ratio between $\hat{P}_1$ when assuming Haldane rather than Eq(3). But, it is worth mentioning here that the loss of support calculated by Morton and our 2:3 ratio are not directly comparable; Morton's calculation was based on real but dependent data points and was estimated by testing the maximum likelihood order against the next maximum one. In our view, it seems that by assuming the Haldane map function when testing gene orders, no real concern should be made about making a wrong decision, the only concern should be the small probability of producing any conclusive one. So that, a useful strategy would be to use this map function as a first step, as it simplifies calculations greatly, and only if an inconclusive result is reached than a second analysis based on a more realistic map function would be needed.

Directing our attention back to our results one can easily see

some obvious limitations that the study suffers from _ N=20, phase known triple backcross, codominant alleles– but another different limitation is the dependence of the study on the criterion mentioned by Ott. In other words, we studied the performance of a certain criterion given a fixed critical level, whereas another interesting extension would be to think about the problem in its reverse order, (i.e) to try to find the critical level which corresponds to a certain required performance. Lathrop et al(1987) started on this road, when testing two orders, by using either of their prescribed – section(2.6)– least favourable or adaptive strategy. Another criticism to our study, that can easily be dealt with, would be that, in some other views, our probability of no conclusion could be further subdivided into probabilities of rejecting one wrong order but not the other, and of no decision at all. This subdivision could perhaps be useful if dealing with more than three loci.

In chapter four and five of the study, we directed our attention to the construction of interval estimates for the unknown recombination fraction(s) $\underline{\theta}$. Various methods based on the large sample properties of the likelihood function were investigated. When the original likelihood is used the empirical joint confidence coefficient for $\underline{\theta}$, for a phase known triple backcross data with sample size N=25, was slighty less than the nominal level of 95% for most of the studied range of $\underline{\theta}$. As for the marginal empirical one, it was quite satisfactory for most of the studied range of $\underline{\theta}$, although some combinations of θ_1 and θ_2 gave quite an unsatisfactory result (θ_1 or θ_2 between [0.07,0.13]). When only few points of the likelihood are supplied the Ott approximation, Morton(1956) or Ott(1985), can be used, the empirical joint and marginal confidence coefficient for this

method were slightly less than that of the original likelihood. To improve on this method we suggested our LSM approximation which gave result very near the original one. The draw back of this last approximation is that, in order to use it in practice, the publication of the pedigree likelihood which is normally supplied in terms of a table of points on a certain grid of θ or $\underline{\theta}$ (usually in step of 0.05 or 0.1) has to be much refined ((i.e) in step of 0.02). It might be worth doing this, especially if we bear in mind that the Ott approximation gave quite an unreliable result as compared to the LSM for some, though not all, combinations of θ_1 and θ_2. In chapter five, we further investigate the performance of either using a high posterior density (HPD) or a central confidence (CCI) Bayesian interval for θ or x as opposed to using the likelihood approach for a double backcross mating with N=25. No clear winner has been found for the studied range of θ or x, nevertheless it was quite clear that the CCI Bayesian interval was unfavourable if either the likelihood or the prior distribution is very much skewed. A natural extension to the work done in those two chapters can be provided by extending the investigation of chapter four to include the phase unknown case and to extend that of chapter five to include the three loci situation.

In chapter six, we have applied the interesting work introduced by Aitchson and Dunsmore(1975) concerning the nature of statistical prediction analysis, to the calculation of the probability of an unborn child being at risk of carrying a genetical disease given his family pedigree, using an example from the genetical literature. This example had, we think, pointed out towards the importance of providing the counselled person with a simple though reliable probability which is

weighted with the whole posterior density function as opposed to a probability which is just the mode of the likelihood one.

APPENDICES

Appendix(1)

The following quotation, from Ott 1985 page 19,20, is given in this appendix to provide the reader with a reference to the DNA recombinant technique which can detect and reveal differences in the DNA sequence between the two homologous chromosomes:

" Differences in DNA sequence can be exhibited as restriction fragment length polymorphisms (RFLPs) in the following way (Botstein et al. 1980; a lucid introduction can also be found in Lange and Boehnke 1983). First, DNA from human lymphocytes is cut into small fragments by DNA restriction enzymes (endonucleases). Such an enzyme recognises a specific sequence in double stranded DNA and cleaves both strands wherever that sequence occurs. The resulting DNA fragments are then separated electrophoretically according to their molecular size. Consider now a particular recognition site on a chromosome of an individual and assume, for example, that the corresponding DNA sequence on the homologous chromosome differs from the recognition site by a base-pair substitution. The altered sequence will then not be cleaved by the restriction enzyme. This and other genotypic differences result in fragments of different lengths. Such RFLPs can be made phenotipically visible as follows. After electrophoresis, the DNA fragments are split into single strands by denaturation, transferred to a solid support, and incubated with radioactive DNA probes (Southern 1975). These probes hybridize only with those fragments that share a homologous DNA sequence with them. Assume now that a probe hybridizes with DNA fragments of differnt lengths that originated from a base-pair substitution at an enzyme recognition site, as postulated above. The resulting RFLP will then show up in autoradiography as two bands (see Botstein

1980, figure 1). The phenotypes of RFLPs are thus quite similar to those of traditional genetic markers detected electrophoretically."

Appendix(2.1)

AIM: To show that the third restriction in (2.3) will be satisfied given a certain map function, $f(\theta)$, (i.e) that:

$$\theta_1+\theta_2-2\theta_1\theta_2 < \theta_{1+2} < \text{Min}[0.5,\theta_1+\theta_2]$$

will be satisfied when taking $\theta_{1+2}= f^{-1}(f(\theta_1)+f(\theta_2))$, where in this appendix we are going to be interested in $f(\theta)$ being equal to either the Haldane, Kosambi or Eq(3). With Haldane we shall have equality at the lower bound for θ_{1+2}.

METHOD: The method is going to depend on the following concept. For a certain function $g(x)$, if the following two conditions are true :

(i) $g(0)\geqslant 0$

(ii) $g'(x)\geqslant 0$ for $0<x<X$

then $g(x)$ is positive for $x\in[0.0,X]$.

But first in order to achieve our aim, we have to divide the above restriction into the following two inequalities:

$$\theta_{1+2} > \theta_1+\theta_2-2\theta_1\theta_2 \qquad \text{(A.2.1.1,a)}$$

$$\theta_{1+2} < \text{Min}[0.5,\theta_1+\theta_2] \qquad \text{(A.2.1.2,a)}$$

Given $f(\theta)$, (A.2.1.1,a) and (A.2.1.2,a) will be equivalent to:

$$f(\theta_1)+f(\theta_2) > f(\theta_1+\theta_2-2\theta_1\theta_2) \qquad \text{(A.2.1.1,b)}$$

$$f(\theta_1)+f(\theta_2) < \text{Min}[f(0.5),f(\theta_1+\theta_2)] \qquad \text{(A.2.1.2,b)}$$

Now let us assume that θ_1 is fixed at a certain value a, where $a\in[0.0,0.5]$, then (A.2.1.1,b) will be satisfied if the following function, $g(\theta_2)$, is positive for $\theta_2\in[0.0,0.5]$, where:

$$g(\theta_2) = f(a)+f(\theta_2)-f(a+(1-2a)\theta_2)$$

Actually this will be easily shown if conditions (i) and (ii) are satisfied; so that for (i):

$$g(0) = f(a)+f(0)-f(a) = 0 \quad \text{, where } f(0)=0 \text{ for any } f(.).$$

As for (ii), recall from chapter one, the basic differential

equation between the map distance, $x=f(\theta)$, and θ which was equal to (formula 1.19):

$\frac{\partial f(\theta)}{\partial \theta} = \frac{1}{1-2c_m(\theta)\theta}$, where $c_m(\theta)$ is the Haldane's marginal coincidence which is equal to 1, 2θ and $(2\theta)^2$ given the Haldane, Kosambi or Eq(3) map function respectively. Now to show that (ii) is satisfied, calculate:

$$\frac{\partial g(\theta_2)}{\partial \theta_2} = 0 + \frac{\partial f(\theta_2)}{\partial \theta_2} - \frac{\partial f(\phi)}{\partial \phi} \times \frac{\partial \phi}{\partial \theta_2} \qquad \text{, where } \phi=a+(1-2a)\theta_2$$

$$= \frac{1}{1-2c_m(\theta_2)\theta_2} - \frac{(1-2a)}{1-2c_m(\phi)\phi} \qquad \text{(A.2.1.3)}$$

Under Haldane, (A.2.1.3) will lead to $g'(\theta_2)=0$ as would be expected. (Recall that $\theta_{1+2} = \theta_1+\theta_2-2\theta_1\theta_2$, given Haldane).

As for Kosambi and from (A.2.1.3), $g'(\theta_2)>0$ if:

$1 - 4\phi^2 - (1-2a)(1-4\theta_2^2) > 0 \quad \leftrightarrow \quad (2\theta_2-1)^2 > 0$

(i.e) for any θ_2 or a.

As for Eq(3) and from (A.2.1.3), $g'(\theta_2)>0$ if:

$1 - 8\phi^3 - (1-2a)(1-8\theta_2^3) > 0 \quad \leftrightarrow \quad (1-2\theta_2)^2[4(1-a)\theta_2+(1+2a)] > 0$

(i.e) if: $\theta_2 > \frac{-(1+2a)}{4(1-a)}$ (A.2.1.4)

So for $a\in[0.0,0.5]$, the R.H.S of (A.2.1.4) will always be negative, (i.e) $g'(\theta_2)>0$ for θ_2 and θ_1 within the feasable region.

Also the second inequality (A.2.1.2,b) will be satisfied, if for $\theta_1+\theta_2>0.5$:

$f(\theta_1)+f(\theta_2) < f(0.5)$ (A.2.1.5,a)

or for $\theta_1+\theta_2<0.5$:

$f(\theta_1)+f(\theta_2) < f(\theta_1+\theta_2)$ (A.2.1.5,b)

But as $f(0.5)=\infty$, for any map function, (A.2.1.5,a) will always be satisfied.

Now let

$h(\theta_2) = f(a+\theta_2) - f(a) - f(\theta_2)$ for $a \in [0.0, 0.5]$ and $\theta_2 \in [0.0, 0.5-a]$

Note that for any map function $h(0)=0$ and:

$$h'(\theta_2) = \frac{-1}{1-2c_m(\theta_2)\theta_2} + \frac{1}{1-2c_m(a+\theta_2)(a+\theta_2)}$$

So that, given Haldane:

$h'(\theta_2) > 0$ if $\theta_2 < (a+\theta_2)$.

And given Kosambi:

$h'(\theta_2) > 0$ if $\theta_2^2 < (a+\theta_2)^2$.

And given Eq(3):

$h'(\theta_2) > 0$ if $\theta_2^3 < (a+\theta_2)^3$.

Which is always true for θ_1 and θ_2 within the feasable region.

Appendix(2.2)

AIM: Generating a data vector $\underline{r}=(r_1\ r_2\ r_3\ r_4)^T$ where:

$$r_1\ r_2\ r_3\ r_4 \sim \text{Mult}(N;\ P_1\ P_2\ P_3\ P_4)$$

METHOD: Generating $\underline{r}$ is equivalent to generating the following three random variables:

$$R_4 \sim \text{bi}(N;\ P_4)$$

$$R_3|r_4 \sim \text{bi}(N-r_4;\ P_3/(1-P_4))$$

$$R_2|r_3r_4 \sim \text{bi}(N-r_3-r_4;\ P_2/(1-P_3-P_4))$$

Generating any bi(N;P) would be then done by making use of the available Nag library routines, which are routine G05EDF and G05EYF.

In general P_1, P_2, P_3 and P_4 would be functions of θ_1, θ_2 and θ_{1+2} as mentioned in the text. But if any map function is assumed then, $\theta_{1+2}=f^{-1}(f(\theta_1)+f(\theta_2))$.

Appendix(2.3)

AIM: To give the full and general proof that for each $P_i(\theta_{ab},\theta_{bc})$ and under a certain map function f(.) then

$P_i(x,y) = P_i(y,x)$ where i= 1,w,4

PROOF: For convenience, let a certain data point $(r_1\ r_2\ r_3\ r_4)$ be denoted by the vector $\underline{r}$ and let the corresponding point $(r_1\ r_3\ r_2\ r_4)$ be denoted by $\underline{r}'$. Where under the true order O_1, both $\underline{r}$ and $\underline{r}'$ comes from the multinomial distribution of (2.5,a). So that

$$p(\underline{r}|x,y)= p(\underline{r}'|y,x) \qquad (A.2.3.1)$$

where $p(\underline{r}|x,y)$ is calculated from the distribution of (2.5,a) when $\theta_1=x$ and $\theta_2=y$.

Also let

S_1= {all $\underline{r}$ such that $\lambda_{12}>2$ and $\lambda_{13}>2$}

S_2= {all $\underline{r}$ such that $\lambda_{12}<-2$ and $\lambda_{23}>2$}

S_3= {all $\underline{r}$ such that $\lambda_{13}<-2$ and $\lambda_{23}<-2$}

Then from the definition of the $P_i(s)$ in (2.7), the following is true

$$P_i(x,y)= \sum_{\underline{r}\in S_i} p(\underline{r}|x,y) \qquad \text{for } i=1,2,3$$

The question that we want to answer now is whether a point $\underline{r}'\in S_i$ if $\underline{r} \in S_i$

Under O_1

$$L_1(x,y;\underline{r}) = \alpha(x,y)^{r_1}\beta(x,y)^{r_2}\gamma(x,y)^{r_3}\delta(x,y)^{r_4} \qquad (A.2.3.2)$$

$$L_1(x,y;\underline{r}') = \alpha(x,y)^{r_1}\beta(x,y)^{r_3}\gamma(x,y)^{r_2}\delta(x,y)^{r_4}$$

where $L_i(x,y;\underline{r})$ is the likelihood function of x and y given the data point $\underline{r}$, order O_i and where α, β, γ and δ are function of x and y as stated in (2.1) and when θ_{1+2} is equal to

$f^{-1}(f(x)+f(y))$. Also notice that as functions of x and y and from (2.1), α, β and γ satisfy <u>in general</u> ((i.e)under any order) the following equalities

$$\alpha(x,y) = \alpha(y,x)$$

$$\beta(x,y) = \gamma(y,x)$$

$$\gamma(x,y) = \beta(y,x)$$

so that

$$\delta(x,y) = \delta(y,x)$$

Therefore

$$L_1(x,y,\underline{r}') = \alpha(y,x)^{r_1}\gamma(y,x)^{r_3}\beta(y,x)^{r_2}\delta(y,x)^{r_4} \qquad (A.2.3.3)$$

By comparing (A.2.3.2) and (A.2.3.3) it is easy to see that

$$\hat{x}_1(\underline{r}) = \hat{y}_1(\underline{r}') \qquad \text{and} \qquad \hat{y}_1(\underline{r}) = \hat{x}_1(\underline{r}'),$$

where $\hat{x}_i(\underline{r})$ and $\hat{y}_i(\underline{r})$ are the MLE of x and y under order O_i and given the data point $\underline{r}$. Therefore

$$\mu_1(\underline{r}) = \mu_1(\underline{r}') \qquad (A.2.3.4)$$

Whereas <u>under O_2</u> and from (2.5,b)

$$L_2(x,y;\underline{r}) = \alpha(x,y)^{r_2}\beta(x,y)^{r_1}\gamma(x,y)^{r_3}\delta(x,y)^{r_4} \qquad (A.2.3.5)$$

And <u>under O_3</u> and from (2.5,c)

$$L_3(x,y;\underline{r}') = \alpha(x,y)^{r_2}\beta(x,y)^{r_3}\gamma(x,y)^{r_1}\delta(x,y)^{r_4}$$

$$= \alpha(y,x)^{r_2}\gamma(y,x)^{r_3}\beta(y,x)^{r_1}\delta(y,x)^{r_4} \qquad (A.2.3.6)$$

By comparing (A.2.3.5) and (A.2.3.6) it is easy to see that

$\hat{x}_2(\underline{r}) = \hat{y}_3(\underline{r}')$ and $\hat{y}_2(\underline{r}) = \hat{x}_3(\underline{r}')$, therefore

$$\mu_2(\underline{r}) = \mu_3(\underline{r}') \qquad (A.2.3.7)$$

Similarly, it is easy to see that

$$\mu_3(\underline{r}) = \mu_2(\underline{r}') \qquad (A.2.3.8)$$

Also it follows from (A.2.3.7) and (A.2.3.8) that if $r_2=r_3$ and therefore $\underline{r} = \underline{r}'$ then $\lambda_{23}(\underline{r}) = \lambda_{23}(\underline{r}') = 1$. Therefore for $\underline{r}=(r_1r_2r_3r_4)$ and $r_2=r_3$, $\underline{r} \notin S_2$ and $\underline{r} \notin S_3$. Also from (A.2.3.4), (A.2.3.7) and (A.2.3.8) it follows that $\lambda_{12}(\underline{r})=\lambda_{13}(\underline{r}')$ and $\lambda_{23}(\underline{r})=\lambda_{32}(\underline{r}')$ which means that if a data point $\underline{r}\epsilon S_1$ then $\underline{r}'\epsilon S_1$ whereas if $\underline{r}\epsilon S_2$ then $\underline{r}'\epsilon S_3$, so that S_1, S_2 and S_3 could be redefined as follows

S_1= {all pair $\underline{r}$ and $\underline{r}'$ such that $\lambda_{12}(\underline{r})>2$ and $\lambda_{13}(\underline{r})>2$}

S_2= {all $\underline{r}$ only such that $\lambda_{12}(\underline{r})<-2$ and $\lambda_{23}(\underline{r})>2$, where $r_2\neq r_3$}

S_3= {all $\underline{r}'$ only such that $\lambda_{13}(\underline{r})<-2$ and $\lambda_{23}(\underline{r})<-2$, where $r_2\neq r_3$}

Then from (A.2.3.1) it follows that

$$P_1(x,y) = P_1(y,x)$$

$$P_2(x,y) = P_3(y,x)$$

$$P_3(x,y) = P_2(y,x)$$

and therefore

$$P_W(x,y) = P_W(y,x)$$

$$P_4(x,y) = P_4(y,x)$$

Appendix(4.1)

AIM: Approximate $\theta = f^{-1}(x)$

by $\theta \simeq g(x)$

where g(x) is a cubic spline function of x.

METHOD:

(i) Data $(\theta_j \; x_j)$ $j=1,2,\ldots,169$

(ii) Regress θ on x using least square method where:

$$\theta_j = C_1N_1(x_j) + C_2N_2(x_j) + \ldots + C_pN_p(x_j) + e_j$$

where:

$e_j = \theta_j - g(x_j)$,

$N_i(x)$ is a normalised cubic B-spline,

C_i are the unknown coefficients to be estimated.

Fitting the function g(x) to θ has been done using the following two NAG routines:

(1)E02BAF: which computes a weighted least square approximation to an arbitrary set of data points by a cubic spline with knots prescribed by the user.

(2)E02BBF: which evaluates the approximating spline at a certain point x supplied by the user.

Appendix(4.2)

AIM: Find the MLE, $\hat{\underline{\Phi}}$ of:

$$Q1(\underline{\Phi}) = a_1\Phi_1^2 + a_2\Phi_2^2 + b_1\Phi_1 + b_2\Phi_2 + c\Phi_1\Phi_2 + d$$

subject to

$$B_1 < \Phi_1 < B_2$$

$$B_1 < \Phi_2 < B_2 \qquad \text{where } B_1=0 \text{ and } B_2=\sqrt{0.5}$$

METHOD:

Firstly let us find the unconstrained maximum $\hat{\underline{\Phi}}_u$. Actually $\hat{\underline{\Phi}}_u$ can be found if:

$$\frac{\partial Q1(\underline{\Phi})}{\partial \Phi_1} = 0 \quad , \quad \frac{\partial Q1(\underline{\Phi})}{\partial \Phi_2} = 0$$

and H, the hessian matrix is negative definite, where

$$H = \left[\begin{array}{cc} \frac{\partial^2 Q1(\underline{\Phi})}{\partial \Phi_1^2} & \frac{\partial^2 Q1(\underline{\Phi})}{\partial \Phi_1 \partial \Phi_2} \\ \frac{\partial^2 Q1(\underline{\Phi})}{\partial \Phi_1 \partial \Phi_2} & \frac{\partial^2 Q1(\underline{\Phi})}{\partial \Phi_2^2} \end{array} \right]_{\underline{\Phi}=\hat{\underline{\Phi}}_u}$$

Under these conditions:

$$\hat{\Phi}_{1u} = \frac{2a_2b_1 - cb_2}{c^2 - 4a_1a_2} \qquad \text{and} \qquad \hat{\Phi}_{2u} = \frac{2a_1b_2 - cb_1}{c^2 - 4a_1a_2}$$

Secondly, if the unconstrained maximum occurs within the feasible region, then $\hat{\underline{\Phi}} = \hat{\underline{\Phi}}_u$, otherwise a unique MLE will occur at one of the boundaries because of the unimodality and concavity of $Q1(\underline{\Phi})$. Actually, because of the concavity of this function (see text, figure(4.3) page 152) any of the boundaries would be tangent to only one contour of $Q1(\underline{\Phi})$ and also at a single point of it. Therefore the boundary which will include the overall maximum, would be the one which is tangent to the nearest contour to the unrestricted maximum provided that the tangent point is within the feasible region. To find the overall maximum, then, the

maximum at each of the possible boundaries will have to be compared. Actually one of the following situation can occur:

(i) If $B_1<\hat{\phi}_{1u}<B_2$ and $B_1<\hat{\phi}_{2u}<B_2$

then $\hat{\underline{\phi}} = \hat{\underline{\phi}}_u$

(ii) If $B_1<\hat{\phi}_{1u}<B_2$ and $\hat{\phi}_{2u}\begin{cases} > B_2 \\ \text{or} \\ < B_1 \end{cases}$

then $\hat{\phi}_1 = \phi_1^*(\hat{\phi}_2)$ and $\hat{\phi}_2 = \begin{cases} B_2 \\ \text{or} \\ B_1 \end{cases}$

(and similarly for $B_1<\hat{\phi}_{2u}<B_2$ and $\hat{\phi}_{1u}>B_2$ or $<B_1$).

(iii) If $\hat{\phi}_{1u}\begin{cases} > B_2 \\ \text{or} \\ < B_1 \end{cases}$ and $\hat{\phi}_{2u}\begin{cases} > B_2 \\ \text{or} \\ < B_1 \end{cases}$

then we will have to compare the following values of the quadratic likelihood to determine the MLE

$Ql(B_2,\phi_2^*(B_2))$ and $Ql(\phi_1^*(B_2),B_2)$

or $Ql(B_2,\phi_2^*(B_2))$ and $Ql(\phi_1^*(B_1),B_1)$, or etc...

where $\phi_1^*(k_2)$ is the MLE of $Ql(\phi_1,k_2)$ subject to $B_1<\phi_1<B_2$, (i.e)

$$\phi_1^*(k_2)\begin{cases} = B_1 & \text{if } k<B_1 \\ = \dfrac{-(b_1+ck_2)}{2a_1} = k & \text{if } B_1<k<B_2 \\ = B_2 & \text{if } k>B_2 \end{cases}$$

Appendix(4.3)

AIM: Given the Haldane map function, find the MLE of:

$$l(\underline{\theta}) \propto r_1 \ln P_1 + r_2 \ln P_2 + r_3 \ln P_3 + r_4 \ln P_4$$

where P_1, P_2, P_3 and P_4 are as described in table(1.1).

METHOD: In general, $r_{34}=r_3+r_4$ and $r_{24}=r_2+r_4$ could be seen as having the following binomial distributions:

$r_{34} \sim bi(n\ P_3+P_4)$ and $r_{24} \sim bi(n\ P_2+P_4)$

where $P_3+P_4= 2\theta_1(1-\theta_1)$

and $P_2+P_4= 2\theta_2(1-\theta_2)$

But given the Haldane map function, r_{34} and r_{24} are mutually independent, which means that $l(\underline{\theta})$ could be rewritten as follows:

$$l(\underline{\theta}) = l_1(\theta_1) + l_2(\theta_2)$$

where $l_1(\theta_1) \propto r_{34}\ln(2\theta_1(1-\theta_1)) + (n-r_{34})\ln(\theta_1^2+(1-\theta_1)^2)$

$l_2(\theta_2) \propto r_{24}\ln(2\theta_2(1-\theta_2)) + (n-r_{24})\ln(\theta_2^2+(1-\theta_2)^2)$

Therefore, the MLE, $\hat{\underline{\theta}}$, could be arrived at by differentiating independently each $l_i(\theta_i)$. From appendix(5.1), we found that the MLE, $\hat{\theta}_i$, of $l_i(\theta_i)$ could have one of the two values:

$$\hat{\theta}_i = \begin{cases} \frac{1}{2} & \text{if } r > \frac{n}{2} \\ \frac{1}{2} - \frac{1}{2n}(n(n-2r))^{0.5} & \text{otherwise} \end{cases}$$

(where $r=r_{34}$ or r_{24} for i=1 or 2 respectively).

Therefore $\hat{\underline{\theta}}$ will be one of the following four possibilities:

(1) $\left[\frac{1}{2} - \frac{1}{2n}(n(n-2r_{34}))^{0.5} ; \frac{1}{2} - \frac{1}{2n}(n(n-2r_{24}))^{0.5}\right]$ if $2r_{34}<n$ and $2r_{24}<n$

(2) $\left[\frac{1}{2} ; \frac{1}{2} - \frac{1}{2n}(n(n-2r_{24}))^{0.5}\right]$ if $2r_{34}>n$ and $2r_{24}<n$

(3) $\left[\frac{1}{2} - \frac{1}{2n}(n(n-2r_{34}))^{0.5} ; \frac{1}{2}\right]$ if $2r_{34}<n$ and $2r_{24}>n$

(4) [0.5 ; 0.5] otherwise

Appendix(5.1)

AIM: Find the MLE of:

$$l(\theta) = r\ln(\phi) + (n-r)\ln(1-\phi)$$

where $\phi = 2\theta(1-\theta)$ and $0<\theta<0.5$

METHOD: Notice that for:

$0 \leqslant \theta \leqslant 0.5 \rightarrow \phi$ is an increasing function of θ and

$$0 \leqslant \phi \leqslant 0.5.$$

This means that we have to maximise the familiar binomial log likelihood as a function of ϕ first, but over the restricted range of $\phi\in[0.0,0.5]$. Thus

$$\hat{\phi} = \begin{cases} \frac{r}{n} & \text{if } \frac{r}{n} \leqslant \frac{1}{2} \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

So that:

$$\hat{\theta} = \begin{cases} \frac{1}{2} - \frac{1}{2n}(n(n-2r))^{0.5} & \text{if } r \leqslant \frac{n}{2} \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

REFERENCES:

Aitchson, J. and Dunsmore, I.R. (1975): Statistical prediction analysis. Cambridge, Cambridge University Press.

Armitage, P. and Berry, G. (1987): Statistical methods in medical research. Oxford Blackwell, 2nd edition.

Bailey, N.T.J (1961): Introduction to the mathematical theory of genetic linkage. Oxford, Clarendon Press.

Boststein, D., White, R.L., Skolnick, M.H. and Davis,R.W. (1980): Construction of a genetic linkage map in man using restiction fragment length polymorphisms. American Journal of Human Genetics, 32, page 314-331.

Carter, T.C and Falconer, D.S. (1951): Stocks for detecting linkage in the mouse and the theory of their design. Journal of Genetics, 50, page 307-323.

Chotai, J. (1984): On the lod score method of linkage analysis. Annals of Human Genetics, 48, page 359-378.

Elston, R.C. and Stewart, J. (1971): A general model for the analysis of pedigree data. Human Heredity, 21, page 523-542.

Gardner, E.J. (1975): Principles of genetics. New York London, 5th edition.

GLIM manual (1987), release 3.77. The Numerical Algorithms Group. Oxford.

Haldane, J.B.S. (1919): The combination of linkage values and the calculation of distances between the loci of linked factors. Journal of Genetics, 8, page 299-309.

Haldane, J.B.S and Smith, C.A.B (1947): A new estimate of the linkage between the genes for haemophilia and colour blindness in man. Annals of Eugenics, 14, page 10-31.

Kosambi, D.D (1944): The estimation of map distances from the

recombination values. Annals of Eugenics, 12, page 172-175.

Lathrop, G.M., Chotai, J., Ott, J. and Lalouel, J.M. (1987): Tests of gene order from three-locus linkage data. American Journal of Human Genetics, 51, page 235-249.

Lathrop, G.M, Lalouel, J.M, Jullier, C. and Ott, J. (1985): Multilocus linkage analysis in human: detection of linkage and estimation of recombination. American Journal of Human Genetics, 37, page 482-498.

Lathrop, G.M, Lalouel, J.M, Jullier, C. and Ott, J. (1984): Strategies for multilocus linkage analysis in humans. Proceedings of the National Academy of Sciences, 81, page 3443-3446.

Maclean, C.J., Morton, N.E. and Lew,R. (1985): Efficiency of lod scores for representing multiple locus linkage data. Genetic Epidemiology, 2, page 145-154.

McCullagh, P. and Nelder, J.A. (1983): Generalized linear models. London New York, Chapman and Hall.

Morton, N.E. (1988): Multipoint mapping and the emperor's clothes. Annals of Human Genetics, 52, page 309-318.

Morton, N.E. (1978): Analysis of crossingover in man. Cytogenetics and Cell Genetics, 22, page 15-36.

Morton, N.E (1956). The detection and estimation of linkage between the genes for Elliptocytosis and the Rh blood types. American Journal of Human Genetics, 9, page 55-75.

Morton, N.E. (1955). Sequential tests for the detection of linkage. American Journal of Human Genetics, 7, page 277-318.

Morton, N.E. and Maclean, C.J. (1984): Multilocus recombination frequencies. Genetics Research Cambridge, 44, page 99-108.

Morton, N.E, Maclean, C.J., Lew, R. and Yee,S. (1986): Multipoint

linkage Analysis. American Journal of Human Genetics, 38, page 868-883.

Nag (1984) FORTRAN library manual. Numerical Algorithms Group, Oxford.

Ott, J. (1985): Analysis of genetic linkage in human families. Baltimore, John Hopkins University Press.

Ott, J. (1977): Counting methods (EM algorithm) in human pedigree analysis. Annals of Human Genetics, 40, page 443-454.

Ott, J. (1974): Estimation of the recombination fraction in human pedigrees - efficient computation of the likelihood for human linkage studies. American Journal of Human Genetics, 26, page 588-597.

Pascoe, L. and Morton, N.E. (1987): The use of map functions in multipoint mapping. American Journal of Human Genetics, , page 174-183.

Rao, D.C., Morton, N.E., Lindsten, J., Hulten, M. and Yee, S. (1977): A mapping function for man. Human Heredity, 27, page 99-104.

Renwick, J.H. (1971): The mapping of human chromosomes. Annual Review of Genetics, 5, page 81-120.

Renwick, J.H. and Bolling, D.R. (1971): An analysis procedure illustrated on a triple linkage for use in prenatal diagnosis of myotonic dystrophy. Journal of Medical Genetics, 4, page 399-406.

Smith, C.A.B. (1986): The development of human linkage analysis. Annals of Human Genetics, 50, page 293-311.

Smith, C.A.B. (1963): Testing for heterogeneity of recombination fractions in human genetics. Annals of Human Genetics, 27, page 175-182.

Wald, A. (1947): Sequential analysis . New York, Wiley.

White, R.L. (1984): Human genetics. The Lancet, December, page 1257-1262.

Yates, J. (1986): Private communications.