SOME STATISTICAL CONSIDERATIONS ON POPULATION STRUCTURE, GENETIC CORRELATION AND HUMAN MULTIPLE BIRTHS

By

RANAJIT CHAKRABORTY

A thesis submitted to the Indian Statistical Institute in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy

Calcutta

1970

ACKNOWLEDGEMENTS

It gives me a great pleasure to express my deep sense of gratitude to Professor C. R. Rao, F.R.S. who introduced me to this field. He was a constant source of inspiration and encouragement to me during the whole course of his supervision of this research work.

I record my sincere regards to him as the Director, Research and Training School, Indian Statistical Institute for providing me with the facilities for carrying out the research at this Institute.

My indebtedness to Professor B. P. Adhikari, Dr. Y. R. Sarma and my colleague Mr. D. C. Rao for numerous inspiring discussions, useful comments and constructive criticisms throughout the progress of the work can be understood by them and me alone. Their kind permissions to include our joint works in this thesis are gratefully acknowledged.

Professor S. R. Das, Professor C. C. Li, Prcfessor R. L. Kirk and Professor P. A. P. Moran were generous enough to spend their valuable time in going through the main results contained in this thesis and offering useful comments. I wish to express my thanks to all of them.

I am thankful to Miss M. Bose who has taken all the patience to proof-read the entire manuscript.

I would like to take this opportunity to thank all the members

of the faculty of the Research and Training School who rendered their help at the various stages of the work.

Finally I thank Mr. P. Nandi whose efforts enabled me to bring the manuscripts to its present form.

RANAJIT CHAKRABORTY

CONTENTS

Page

CHAPTER IV:

CHAPTER 0

ABOUT THE THESIS

0.1 INTRODUCTION

Since the rediscovery of Mendel's works towards the beginning of this century the geneticists did not confine their studies only at familial level. The study of genetic entities at population level also became equally important for understanding the mechanism of inheritance. This branch of understanding the mechanics of heredity, known as Population Genetics, has by now become so well known that a discussion of any part of it does not need any general introduction. In view of this fact, we instead spend some time to get into the problems discussed in this thesis.

Basically we study three problems in this thesis. The first problem is that about population structure. Though population as a whole is of interest to the population geneticists, his knowledge is incomplete without an idea about the constituents of the population. We, in fact, study the structure of populations which are under genetic equilibrium, the mating models which are likely to be in operation in such an equilibrium population and further some statistical aspects of estimation of gene frequencies in two common polymorphic systems, namely ABO and MNSs blood group systems. With special reference to ABO blood group system it is shown that some population structures are not distinguishable from others under some conditions.

The second problem is the study of genetic correlations. The importance of the role of genetic correlations in understanding the inheritance of genetic characters, especially the quantitative ones, is by now beyond doubt. Though much has been done in this field of population genetics most of the results were applicable to random mating population only. The population structure prescribed by Wright (1921), widely known as Model II population structure, received relatively less attention although Li (1955) and Kempthorne (1957) have considered them in some details.

Although the first two problems have a common base, in the sense that both stem out from the basic problem of understanding the population constituents and their interrelations, the third problem stands out to be a discrete one in this respect. This is about the construction of some mathematical models for human multiple births. In this field also there had been many contributions to know the exact factors influencing the human multiple births but it is felt that these biological findings were not properly injected in the construction of the models present in the literature (Das, 1953-56; Bulmer, 1958; Allen, 1960).

0.2 POPULATION STRUCTURE

One of the most important concept in this area is that of genetic equilibrium. But the term is more often than not misappropriated. Many text books interpret genetic equilibrium as equilibrium at gene level. This far is quite sound. But when one says, 'one characteristic feature of a random mating population is that in it the gene frequencies are kept constant generation after generation barring the influences of external forces like mutation, migration and selection'. It is to be remembered that this is no criterion of random mating. It is a fact that for any closed population where forces like mutation, migration and selection are inoperative gene frequencies remain same generation after generation irrespective of the mating system practiced there. Apparently this was the contention of J. B. S. Haldane, who called it as 'Gene Pool Theorem'. Therefore, to characterise any mating system, it is not sufficient to consider equilibrium at gene level. We, in section 1.1 consider only genotypic equilibrium and derive the necessary and sufficient conditions for the maintenance of such equilibrium. It turns out to be that random mating, among many other mating systems, in one mating scheme which keeps a population under genotypic equilibrium. Li (1955) studied such conditions for an autosomal trait explained by two codominant alleles at a locus. We, in section 1.2, extend this to K-allelic case and also prescribe the conditions for a sex-lined character. The main results indicated in this section are presented also in Chakraborty (1970c). The analysis of the K-allelic case shows that the class of mating structures which preserve a population at genotypic equilibrium is too wide to be studied in great details. We, therefore, restrict ourselves to a smaller class, which is known as Model II population structure.

Though, in 1921 Sewall Wright prescribed this structure of an equilibrium population there had been many attempts thereafter to interpret the parameters used therein. The parameter, F, when different from zero signifies the deviation from random mating. A value of F = 1 characterises a mating system where there is complete fixation of the genes. It is for this reason F is sometimes referred to as coefficient of fixation. Although most frequently F is interpreted as inbreeding coefficient, through a discussion of the various interpretations of this F coefficient in section 1.2 our contention is to show that among many other forces which may cause deviations from random mating, consanguinity or inbreeding is only one factor.

But both Model I (random mating) and Model II population structure suffer from some mathematical and statistical loopholes. It is shown in Chapter IV (section 4.1.3) that under some conditions a mixture of random mating isolates cannot be distinguished from a homogeneous random mating population and hence,if, without knowing exact composition of the population, one blindly applies the random mating formulae to compute the gene frequencies of the population, he gets solutions to some mathematical equations but with no physical significance attached to them. Model II population structure on the other hand, provides little scope for statistical estimation of the parameters. As is shown in section 4.1.4, in case of ABO blood groups, only in a very restricted set up Schull's (1965) explicit expressions give rise to admissible value of gene frequencies and the F-coefficient. On the other hand, even if such solutions are admissible there is no guarantee about the reliability of such estimates. Sometimes standard error of the estimates are so high that even an appreciable amount of inbreeding cannot be detected in large samples. The problem of detection of this F coefficient from ABO blood group data is discussed from another angle in section 4.1.5. Minimum sample size is computed to detect a prescribed level of F with given precision (in terms of the level and power of the test procedure).

With these limitations of the mating structures of Model I and Model II, it was felt to design a scheme wherein the analysis is plausible for mating structures which do not come in either model. Such a mating model is termed as Restricted Random Mating (R.R.M.). The terminology, itself, suggests that the assumption of random mating is not dispensed with in full. In fact the mating at the gametic level is assumed to be at random. Invoking Haldane's 'Gene Pool Theorem', it is obvious to note that the gene frequencies are kept constant in a population where the mating in vogue is R.R.M. In Chapter II we discuss the problem of estimation of ABO gene frequencies under this set-up, evolve a goodness of fit statistic and derive its asymptotic sampling distribution. Such analysis is of great use especially when there are dominance relationships among the alleles present at the locus of interest. The analysis is also illustrated through data on S-factor of MNS blood group system (section 2.3).

0.3 GENETIC CORRELATIONS

The latter half of the Chapter I is devoted to the study of genetic correlations in a population whose structure is slightly general than Model I. Under this Model II structure, the results of Li (1954) are generalized. Li has derived the correlation between the parents and the offspring when both parents and s (>1) offspring are considered. He termed this correlation as 'parent offspring correlation'. Li's results were valid only for random mating populations and he considered the case with two codominant alleles at an autosomal locus. In sections 1.4 and 1.5 we derive the parent offspring correlation for autosomal as well as sex-linked characters wherein the alleles do or do not have any dominance relationship between them. It is seen that in the absence of any dominance relationship between the alleles, the parent offspring correlation depends solely upon F, the coefficient of departure from random mating but as soon as the characters are recessive in nature, the parent offspring correlations also involve the gene frequencies. Generalizations are seeked in K-allelic (section 1.6) case also for which we invoke the weighting system developed by Stanton. This is necessary because the superficial weighting scheme followed otherwise presents obvious anomalies (Chakraborty, 1970a, 1970b, 1970c).

In section 1.7, we derive some important correlations between relatives other than parents and offspring. It is to be remembered that the environmental effect is altogether ignored while deriving the different correlations in this chapter. The last section of chapter I (section 1.8) advocates another important role of parent offspring correlation. Parent offspring correlation is used here to estimate the coefficient of departure.from random mating. Though for such estimation of F, family materials are necessary but the advantage of this estimate over the existing ones is also indicated in this section. The materials of this section is also found in Chakraborty (1970d).

0.4 HUMAN MULTIPLE BIRTHS

The biological mechanism of human multiple births is, by now, sufficiently explained. Though not the exact nature, but the hereditary nature of dizygotic twinning tendency is known and further we know that twining rate changes with parity of birth as well as the age of the mother. Again the behaviours of monozygotic and dizygotic twinning rate are not the same in, at least, these two respects. In Chapter III, first we give an account of these biological findings. Later on in section 3.4 we construct a probabilistic model for these multiple births which is in fact a generalization of Bulmer's Model (1958) taking into account the fact that hereditary property of dizygotic twinning was not incorporated in Bulmer's Model. In section 3.5 we present a stochastic model where the intensities of the two basic processes (namely, scission of a zygote and the release of extra eggs) are assumed to be time dependent. In the last section of this chapter the intensity of the release of extra ova is assumed to be dependent also on the age of the mother. One plausible relationship is also indicated through an empirical study. The models constructed in this chapter thus take the biological findings also into account and present neat expressions without introducing any parameter with no physical interpretation as seen in Das (1953-56) and Allen (1960).

In the last chapter of the thesis, in section 4.1.2, a comparison of the different method of estimating the ABO gene frequencies (under random mating population structure) is tried out. DeGroot (1956) had tried one such comparison where instead of considering all of the estimates at a time, he took one at a time and compared the efficiencies measuring them by reciprocal of the variances. However, we take the generalized variance as the criterion and draw the comparison, The results those we obtain are same as those of DeGreet (1956) and Sukhatme (1942). In section 4.2 the MNSs system is discussed. In this system the gene count method of estimating the chromosome frequencies is slightly modified for the use of desk calculators. The loss of efficiency (or, the increase in standard errors) are not appreciable. These two results are also found in Chakraborty (1970e, 1970f).

CHAPTER I

GENETIC CORRELATIONS IN A GENERAL EQUILIBRIUM POPULATION WITH ONE LOCUS SEGREGATING

1.0 INTRODUCTION

At the dawn of this century geneticists rediscovered and established the Mendelian principles from studies at the familial level while searching for the mechanisms of heredity. They confined their investigations to the alikeness or unlikeness between specified parents and their offspring. Population genetics, on the other hand, is concerned with the statistical consequences of Mendelism in a 'group' of families or individuals; it understands the hereditary phenomenon on a population level. A population geneticist goes on to investigate the proportions of purple- and white-flowered plants in a given region, the frequencies of the various types of crosses in such a regional population, the proportions of the various kinds of plants from each type of cross, and the genetic structure of one generation as compared with that of the next under various circumstances presuming the mechanism of heredity to be what Mendelian genetics has described.

The life of an individual is limited in length of time, and barring mutation his genetic makeup is fixed throughout his life. In contrast, a population is practically immortal, may be large or small in bulkiness, may be spread over a wide or limited geographical area, and may change in genetic composition from generation to generation, suddenly or gradually. The study of population genetics is thus inevitably related with the understanding of its genetic makeup. In this chapter we shall first deal with some principles and laws which emerge while making an attempt to study the Mendelian consequences in a population and with this background we shall study the genetic relationship between the parents and offspring which, needless to say, plays a crucial role in investigating the mechanism of heredity.

1.1 EQUILIBRIUM CONDITIONS

The first and foremost important concept of population genetics which goes as the most important landmark in the subject is genetic equilibrium. By 'equilibrium' we mean that there is no change in genotypic proportions in a population from generation to generation. This implies no changes in gene frequencies either. There are many possible types of equilibrium conditions of which we study the most important ones only. The particular equilibrium condition under random mating (a mating system in the case of bisexual organisms where any one individual of one sex is equally likely to mate with any individual of the opposite sex) is known as HARDY-WEINBERG LAW because it was discovered independently by Hardy and by Weinberg in the same year, 1908 (see for example Stern, 1943). This law may be formulated as follows:

Consider a large panmictic population wherein there are only two alleles (A, a) at a locus with relative frequencies p and q $(p + q = 1)$. If the proportions of the three genotypes AA, Aa and aa with respect to this pair of genes in the population are p^2, $2pq$ and q^2 respectively, the genotypic proportions in the next generation will be the same as those in the preceding generation. The population $(p^2, 2pq, q^2)$ is then said to be in equilibrium in Hardy-Weinberg's sense under the system of random mating.

A nice property of such an equilibrium was shown by Wentworth and Remick (1916) wherein they proved that equilibrium in Hardy-Weinberg's sense is reached after a single generation of random mating regardless of the initial composition of the population.

The direct extension of the Hardy-Weinberg Law to the case of multiple alleles was first made by Weinberg (1909; see Stern, 1943), the general formulation of which can be written as follows:

In a large panmictic population in which the frequency of the allele A_i is q_i, the proportions of the various genotypes in an equilibrium condition are given by the coefficients of the A's in the expression

$$\left[\sum_i q_i A_i\right]^2 = \sum_i q_i^2 A_i A_i + 2 \sum_{i<j} q_i q_j A_i A_j \qquad (1.1.1)$$

where $\sum_i q_i = 1$ ($i = 1, 2, \ldots, k$). We shall refer to this population in our subsequent discussions as MODEL I. It may be noted that if an initial population is not in an equilibrium state, the condition (1.1.1) will be immediately established after one single generation of random mating, just as in the case with two alleles.

A more general result is known in the two allelic case which can be stated as follows:

THEOREM 1.1.1 (Li, 1955) A population will be in equilibrium with respect to an autosomal locus with two alleles A and a, if and only if the Aa x Aa matings are twice as frequent as those between the two different homozygotes (AA x aa and aa x AA).

For a proof of this result one can refer to Li (1955). Note that this relation is independent of gene frequencies or amount of consanguinity. Also, it is clear that such a relation is true for panmictic populations. This property of an equilibrium population was first noted by Fisher (1918, pp. 410-11) whose argument is especially simple. Of the six possible types of matings in the population, four types (AA x AA, aa x aa, AA x Aa, aa x Aa) produce offspring of the same genotypic proportions as their parents. On the contrary, in AA x aa matings the two homozygous parents are replaced by heterozygotes in the next generation whereas in Aa x Aa matings only half of the offspring regain the homozygous condition. Hence the theorem. It is to be noted here that the complications of the selection pressure is altogether ignored for this purpose. All the genotypes are assumed to have the same fitness coefficients.

Analogous conditions can also be studied in the case of sex-linked characters. As usual, one may take the homogametic XX individuals as females and the heterogametic XY (or XO) as males, where X denotes the sex chromosome. Then, with respect to a X-linked locus with two alleles A and a, in case of females we have genotypes AA, Aa and aa whereas the males can be of genotype A or a. With this, the mating matrix can be represented by TABLE 1.1.1.

TABLE 1.1.1

Mating frequencies with sex-linked genes

Females	Males		Total
	A	a	
AA	u_{21}	u_{20}	Q_2
Aa	u_{11}	u_{10}	Q_1
aa	u_{01}	u_{00}	Q_0
Total	P_1	P_0	1

With this, it is easy to see that the zygotic proportions of the females in the next generation are given by

$$Q_2' = u_{21} + \frac{u_{11}}{2}$$

$$Q_1' = u_{20} + \frac{u_{11}}{2} + \frac{u_{10}}{2} + u_{01}$$

$$Q_0' = \frac{u_{10}}{2} + u_{00}$$

which in turn give the necessary and sufficient conditions for genetic equilibrium as $u_{11} = 2\,u_{20}$ and $u_{10} = 2\,u_{01}$. It can be verified that these conditions also imply that the male frequencies are kept constant in the next generation. Thus we have :

THEOREM 1.1.2 (Chakraborty, 1970) A population will be in equilibrium with respect to a X-linked locus with two alleles A and a if and only if frequency of Aa x A mating = 2 x frequency of AA x a mating and frequency of Aa x a mating = 2 x frequency of aa x A mating.

Similar results can also be obtained with more than two alleles at a locus. But as the number of alleles increases, the number of conditions also increases enormously. With three alleles A_1, A_2, A_3 at an autosomal locus, it may be seen that there are as many as six equations which together form a set of necessary and sufficient conditions for genetic equilibrium. The mating frequencies in such a situation can be designated by TABLE 1.1.2.

TABLE 1.1.2

Mating frequencies with 3 alleles at an autosomal locus

Mates	A_1A_1	A_1A_2	A_1A_3	A_2A_2	A_2A_3	A_3A_3	Totals
A_1A_1	u_{11}	u_{12}	u_{13}	u_{14}	u_{15}	u_{16}	U_1
A_1A_2	u_{12}	u_{22}	u_{23}	u_{24}	u_{25}	u_{26}	U_2
A_1A_3	u_{13}	u_{23}	u_{33}	u_{34}	u_{35}	u_{36}	U_3
A_2A_2	u_{14}	u_{24}	u_{34}	u_{44}	u_{45}	u_{46}	U_4
A_2A_3	u_{15}	u_{25}	u_{35}	u_{45}	u_{55}	u_{56}	U_5
A_3A_3	u_{16}	u_{26}	u_{36}	u_{46}	u_{56}	u_{66}	U_6
Totals	U_1	U_2	U_3	U_4	U_5	U_6	1

From this, one easily gets the necessary and sufficient conditions for equilibrium as

$$\left.\begin{array}{l} u_{22} = 4\,u_{14},\ u_{23} = 2\,u_{15} \\ u_{55} = 4\,u_{46},\ u_{25} = 2\,u_{34} \\ u_{33} = 4\,u_{16},\ u_{35} = 2\,u_{26} \end{array}\right\} \quad \ldots \qquad (1.1.2)$$

For a sex-linked locus with three alleles the number of conditions is even more. One can easily see that in such a case there are as many as eight equations which form a set of necessary and sufficient conditions.

Fisher's argument for the validity of Theorem 1.1.1 can easily be extended to form a set of necessary and sufficient conditions for K alleles ($K \geq 2$) at an autosomal locus. In such a case we have $\frac{K(K - 1)}{2}$ number of equations of the form

$$\left.\begin{array}{l}
\text{Freq. of } A_iA_j \times A_iA_j \text{ matings} = 4(\text{Freq. of } A_iA_i \times A_jA_j \text{ matings}) \\
\qquad \text{for all } i < j \text{ ; } i, j = 1, 2, \ldots, k \\
\text{and } K(K - 1)(K - 2)/2 \text{ equations of the form} \\
\text{Freq. of } A_iA_j \times A_iA_k \text{ matings} = 2 \times \text{Freq. of } A_iA_i \times A_jA_k \text{ mating} \\
\qquad \text{for all } i, j, k = 1, 2, \ldots, K \text{ such that } j < k \\
\qquad \text{and } i \neq j \neq k.
\end{array}\right\} (1.1.3)$$

Note that total number of equations in such a set is $K(K - 1)^2/2$. Thus putting $K = 3$, we have 6 equations, for four allelic case one has 18 equations and so on.

Because of these many equations, it is very difficult to study the populations under such equilibrium conditions.

Wright studied yet another type of equilibrium situation in his classic study of 1949 (Wright, 1949). Though his equilibrium conditions are not as general as (1.1.3) but certainly it is a generalization over the panmictic equilibrium model (Model I). In his model he pointed out that when the gametes are not uniting entirely at random but are correlated, there will be relatively more homozygous individuals in the population than in Model I. If the correlation coefficient between the uniting gametes is F, the population, in equilibrium state, will consist of :

$$\begin{aligned}(1 - F) \left[\sum_i q_i A_i \right]^2 + F \sum_i q_i A_i A_i \\ = \sum_i \left[(1 - F) q_i^2 + F q_i \right] A_i A_i \\ + 2(1 - F) \sum_{i < j} q_i q_j A_i A_j\end{aligned} \qquad (1.1.4)$$

We shall refer to this population as MODEL II. When $F = 0$, it reduces to model I and thus model II is a generalization of model I. It should be remarked that the parameter F, measuring the degree of association between the uniting gametes, is entirely independent of the gene frequencies, q_i's. The later tells us what proportion of each allele there is in the population while the former measures the extent of the association between pairs of the alleles. In (1.1.4), the F is assumed to have values in the closed interval $[0, 1]$. In the next section we shall study in details the various interpretations of this parameter, F. That the equilibrium condition of Wright (as given by 1.1.4) satisfies the general equilibrium theorems 1.1.1 and 1.1.2 follows from Yasuda (1968).

Yasuda based his results on what is known as Wahlund's principle. Suppose that a population is divided into many endogamous panmictic smaller populations (isolates) restricted by geographic.., racial, religious, social and economic barriers. Let w_i ($\Sigma w_i = 1$) be the relative size of the i^{th} isolate. If a genetic system consists of two alleles A and a with frequencies p_i and q_i in the i^{th} isolate, respectively, then the frequency p of gene A is $p = \Sigma p_i w_i$ and its variance σ^2 in the total population is $\Sigma(p_i - p)^2 w_i = \Sigma p_i^2 w_i - p^2$, where the summation is taken over all isolates. Since the genotypio frequencies of AA, Aa and aa in the total populations are $\Sigma p_i^2 w_i$, $2 \Sigma p_i q_i w_i$ and $\Sigma q_i^2 w_i$, respectively, the subdivision results in increasing homozygosity by an amount equal to the gene frequency variance σ^2. Comparison of this result with that of Wright (equation 1.1.4, taking $K = 2$) leads to $\sigma^2 = p(1 - p)F$.

In case of a continuous model the gene frequency and its variance in the population can be expressed by Labesgue-Stieltjes integrals

$$p = \int p_w \, dw \quad \text{and} \quad \sigma^2 = \int p_w^2 \, dw - p^2,$$

where sums are taken for the discrete model and integrals for the continuous model. Note that, thus, the first moment of the isolate-distribution gives the gene frequency and the second moment the genotype frequency. The third and the fourth moments give the mating type frequencies at a sex-linked and at an autosomal locus, respectively, since three and four genes are concerned in each gene combination.

Let us consider a locus with two alleles A and a whose frequencies are p and q, respectively, in a subdivided population with a non-randomness coefficient F. Suppose that the difference between gene frequency of an isolate, p_w, and the population, p, is Δp_w, whose k^{th} moment is expressed by m_k :

$$m_k = \int (\Delta p_w)^k . \, dw = \int (p_w - p)^k \, dw,$$

where the integrals are understood in the Lebesgue-Stieltjes sense.

For the population moment, as Yasuda (1968) shows, M_a,

$$\begin{aligned} M_a &= \int p_w^a . \, dw = \int (p + \Delta p_w)^a \, dw \\ &= \sum_{r=0}^{a} \left[\binom{a}{r} p^{a-r} \int (\Delta p_w)^r \, dw \right] \\ &= \sum_{r=0}^{a} \binom{a}{r} p^{a-r} m_r ; \end{aligned}$$

or $M_a = p^a + \frac{a(a-1)}{2} p^{a-1}(1 - p). F + O(m_3)$,

where $O(x)$ stands for any function which is at most of order x. In the above expression, if the cubic and higher powers of Δp_w are negligible, the term $O(m_3)$ can be ignored. For example

$$\begin{aligned} M_1 &\doteq p, \\ M_2 &\doteq p^2 + p(1 - p) F, \\ M_3 &\doteq p^3 + 3p^2(1 - p) F, \\ M_4 &\doteq p^4 + 6p^3(1 - p) F. \end{aligned} \qquad (1.1.5)$$

With this it is straightforward to evaluate the mating type frequencies in the case of two alleles at an autosomal and a sex-linked locus (for autosomes, reciprocal crosses are grouped together). To illustrate, consider the intercross Aa x Aa and its relative frequency u_{11}. In an isolate, the proportion of this mating type is $4p_w^{\,2}(1 - p_w)^2\,dw$, so that

$$\begin{aligned} u_{11} &= \int 4p_w^{\,2}(1 - p_w)^2\,dw \\ &= 4\,M_2 - 8\,M_3 + 4\,M_4 \\ &\doteq 4p^2q^2 + 4pq(1 - 6pq).F \end{aligned} \qquad (1.1.6)$$

The mating types and their relative frequencies, thus computed, are shown in TABLE 1.1.3 for an autosomal locus and in TABLE 1.1.4 for sex-linked locus.

TABLE 1.1.3

Frequency of mating types (two alleles at an autosomal locus)

Mating type	Frequency
AA x AA	$p^4 + 6p^3\,q.F$
AA x Aa	$4p^3q + 12p^2q(1 - 2p).\,F$
Aa x Aa	$4p^2q^2 + 4pq(1 - 6pq).\,F$
AA x aa	$2p^2q^2 + 2pq(1 - 6pq).\,F$
Aa x aa	$4pq^3 + 12pq^2(1 - 2q).\,F$
aa x aa	$q^4 + 6p^3\,q.F$
Total	1

TABLE 1.1.4

Frequency of mating types (two alleles at a sex-linked locus)

Mating type	Frequency
AA x A	$p^3 + 3p^2 q.F$
Aa x A	$2p^2q + 2pq(1 - 3p). F$
AA x a	$p^2q + pq(1 - 3p). F$
Aa x a	$2pq^2 + pq(1 - 3p). F$
aa x A	$pq^2 + pq(1 - 3p). F$
aa x a	$q^3 + 3pq^2. F$
Total	1

Once these two tables are ready, it is easy to see that these mating frequencies satisfy the general equilibrium conditions as dictated by Theorem 1.1.1 and Theorem 1.1.2. Thus, it is proved that Equilibrium in Wright's sense (prescribed by Model II of equation 1.1.4) is a particular case of the general genetic equilibrium but Model II is a generalization of Model I.

1.2 MODEL II : INTERPRETATION OF WRIGHT'S F PARAMETER

We have already seen that Model II describes the genotypic frequencies in an equilibrium population in terms of the gene-frequencies

q_i's and the parameter F denoting the so-called fixation index. As we have seen in section 1.1, the value of F describes the deviation from Hardy-Weinberg proportions (as given by equation (1.1.1)) due to the net effect of all of the forces acting on the genes and genotypes at the locus under consideration. If inbreeding in a constant amount, f, per generation, is the only force acting on the population, then F will also be a constant, equal to f, the so-called "coefficient of inbreeding". In reality, however, there may be (1) inbreeding due solely to the finite size of the population (f_N) which may vary from one generation to the next as the effective size of the population varies, (2) inbreeding due to some regular pattern of consanguinity (f_c) and, (3) inbreeding due to positive assortative mating (homogamy) which may vary among genotypes or between sexes. In addition, F will also be affected by selection (e.g., differential viability or fertility), mutation, gene flow, and random genetic drift (Jain and Workman, 1967).

Whatever may be net operative force on the locus, several interpretations of this F-parameter have been attempted (Wright, 1921; Bernstein, 1930; Malécot, 1948) and the same conclusion was reached with respect to zygote frequencies in terms of gene frequencies and the inbreeding coefficient (i) This F-parameter can be understood as a measure of non-randomness that also describes zygote frequencies, the correlation between uniting gametes (Wright, 1921). Wright's this interpretation is already indicated in the previous section. But it may be worthwhile to review the works of Malécot (1948) which resulted in essentially the same formulas as that of Wright since it dictates the crucial ideas untrammeled by unnecessary assumptions.

The basic notion in Malécot's presentation is that two genes in the population may be alike for two entirely exclusive reasons :

DEFINITION 1.2.1 : Two genes are said to be identical by descent if they are replica of the same gene possessed by some ancestor.

Thus they are alike because they are copies arising in the reproductive process of one gene occurring previously in the ancestry, or one is a copy of the other.

DEFINITION 1.2.2 : Two genes chosen one from each of two unrelated individuals are said to be alike in state if they are found to be in the same state.

That is, they may be alike in the sense of being both A: for example, because two genes are drawn at random from the population and both happen to be A. If the gene frequency for A is p, then the probability of two randomly drawn genes being both A is p^2. Sometimes they are also called identical by state.

For instance, let us take the population p^2 AA + 2pq Aa + q^2 aa

(Model I). We draw two individuals at random from it and then consider a gene at random from each individual. The probability of the genes being alike in state is equal to $p^2 + q^2$.

With this background we are now in a position to quantify the degree of relationship between two individuals precisely. Malecot uses the term "coefficient de parente" for which exact English term is not available yet. Instead Wright (1922) used the term "coefficient of relationship". Nevertheless, for avoiding confusion, it is to be stressed that they are not the same thing. In fact under panmixia "coefficient of relationship denotes a quantity which is twice Malecot's 'coefficient de parente'". Kempthorne (1957) translated Malecot's "coefficient de parente" as "coefficient of parentage". This can be defined as follows :

<u>DEFINITION 1.2.3</u> : Consider two individuals X and Y with genotypes ab and cd (where a, b, c and d may be A or a independently). Then r_{XY} is defined to be the probability that a random gene from X is identical by descent with a random gene from Y.

In another language, if we use $P(a = c)$, say, to denote the probability that genes a and c are identical by descent then

$$r_{XY} = \frac{1}{4}\left[P(a = c) + P(a = d) + P(b = c) + P(b = d)\right] \qquad (1.2.1)$$

Now we define the coefficient of inbreeding of an indiuidual as :

DEFINITION 1.2.4 : Coefficient of inbreeding of a diploid individual (with respect to a fixed locus) is the probability that the two genes possessed by that individual at that locus are identical by descent.

For instance ., inbreeding coefficient of an individual X may be denoted by F_X. If X has the genotype ab at a locus, then by definition

$$F_X = P(a = b).$$

Comparing X with itself, it is easy to find that

$$r_{XX} = \frac{1}{2}(1 + F_X). \tag{1.2.2}$$

Thus we have the coefficient of parentage of X with itself equals unity plus the coefficient of inbreeding of X, whole divided by 2.

Wright's formulation for computing the coefficient of parentage is as follows :

Firstly, observe that the only contributions to the coefficient of parentage of two individuals X and Y arise from lines of ancestry leading from X and Y to common ancestors. If we designate by Z a common ancester which is n_X steps above X and n_Y steps above Y, then it is clear that the only contribution to the coefficient of parentage arising from the chain of relationship from X to Z to Y is equal to

$$(\frac{1}{2})^{n_X + n_Y} \cdot r_{ZZ}$$

or

$$(\frac{1}{2})^{n_X + n_Y + 1} (1 + F_Z) \quad \text{(using equation (1.2.2))}.$$

To get the total coefficient of parentage we merely consider all possible distinct chains of X - Z - Y relationships and add up the contributions. Thus

$$r_{XY} = \sum (1/2)^{n_X + n_Y + 1} (1 + F_Z) \qquad (1.2.3)$$

summation being taken over all distinct chains of X - Z - Y relationships (Wright, 1921).

It may be of interest to mention here a few notes on these two coefficients made by Kempthorne (1957). (i) These coefficients are probabilities which bear no relation whatsoever to the gene effects, Of course, it is true that these coefficients do appear into the correlations between two relatives with respect to characters which are either quantitative or transformed into so using the dichotomy of the qualitative nature of the character. Kempthorne goes further to term such usage of these coefficients as "appendage" without having any real basis. (ii) If one interprets F as the correlation between uniting gametes (as Wright did), some quantative attribute is to be assigned to each gamete in such a case. In the two allelic cases it poses no real trouble since one can always construct a two-by-two table, merely insert the frequencies in each cell and attach numbers 1 and 0 to A and a, respectively, and compute the product moment correlation between the numbers for the gametes of the mating partners. However, these formulation clearly takes one into deep trouble in case he has to deal with more than two alleles at a locus.

The above formulation is made in such a way that the coefficients of inbreeding and parentage do not depend upon the number of alleles present at that locus. Thus, if we start with the population $(\Sigma\, q_iA_i)^2$ and choose individuals at random to enter the pedigree, then it is clear enough that the probability that any given gene is A_i is q_i. If F is the probability that two genes are identical by descent, the probability that they are both A_i is Fq_i. The probability that they are not identical by descent is (1 - F), and the probability that two ordered original genes are A_i and A_j is q_iq_j. Hence we can state that the population which would result from inbreeding the population $(\Sigma\, q_i\ A_i)^2$ (as dictated by Model I) to an extent measured by F has the array similar to Model II. To be specific, in the case with two alleles A and a, the array is

$$[Fp + (1 - F)\ p^2]\ AA + [2\ (1 - F)\ pq]\ Aa + [Fq + (1 - F)\ q^2]\ aa$$

where p and q are the frequencies of A- and a-alleles. This may, alternatively, be written as

$$[p - (1 - F)\, pq]\, AA + [2(1 - F)\, pq]\, Aa + [q - (1 - F)\, pq]\, aa$$

or as

$$[p^2 + Fpq]\, AA + [2(1 - F)\, pq]\, Aa + [q^2 + Fpq]\, aa$$

indicating that homozygotic classes are each increased by Fpq which is one half of the loss of heterozygotes resulting from inbreeding.

Another attempt at interpreting the same F-parameter is Bernstein's α-coefficient (Bernstein, 1930). The formulation goes as follows :

Let $P[A/A] = P$ denote the conditional probability of uniting with an A gamete when the given gamete is known to be A. Similarly, $P[a/a] = Q$ can also be defined. Now, since the probability that any gamete given at random should be A is p, and that be a is q (assuming the allelic frequencies to be p and q for A and a, respectively), the probability that any zygote be AA is $p \times P$ or that it be aa is $q \times Q$. Further, the probability that any zygote be Aa or aA is $2p(1 - P)$ or $2q(1 - Q)$. Equality of these last two expressions yield the equation

$$\frac{1 - P}{q} = \frac{1 - Q}{p} = 1 - \alpha$$

which defines Bernstein's α-coefficient $(0 \leq \alpha \leq 1)$. It can easily be verified that Bernstein's α-coefficient is the same as Wright's F-coefficient (for verification see Li, 1955).

Besides these, F-parameter can also measure the degree of differentiation in subdivided populations and describe mating type frequencies (Wahlund's effect; Wahlund 1928). But with this interpretation the situation remains no longer identical with that of inbreeding in the multi-allelic case since in the latter case (inbreeding), all the heterozygote frequencies are decreased to the same extent whereas in the former (population subdivision) a heterozygote frequency may be decreased or increased, or remains the same as that of random mating population without subdivision, as the covariance of the frequencies of the alleles A_i and A_j may be negative, positive, or zero (Li, 1969). Also, no correlation coefficient can be calculated for the case of population subdivision, as no natural numerical values can be assigned to the alleles A_i.

Li (1955) indicated several other interpretations of this F coefficient by relating it with indices of combination of panmixia with selfing and panmixia with sib-mating. But, for human genetics those relationships do not carry much importance although for plant breeding purposes such formulations can be fruitfully employed.

In the sequel we call the F-parameter as the coefficient of non-randomness since basically it signifies the actual departure from random mating. Note that $0 \leq F \leq 1$. Since Model II with $F = 1$ describes a population where all the genes are fixed (all individuals are genetically homozygotes), F is sometimes referred as the fixation index also.

1.3 GENETIC CORRELATIONS UNDER PANMIXIA

In the earlier two sections we have discussed the nature of genetic equilibrium and presented the structure of a population which preserves genetic equilibrium with respect to one segregating locus. Having done so, we now proceed to study a powerful tool to study the mechanism of inheritance of characters which are metric in nature or may be made so by using the dichotomy of its qualitative nature. This tool is nothing but genetic correlation or correlations of genotypes among relatives. Extensive study has been done on this subject most of which are for random mating populations. In this section we are going to give a brief account of the important ideas on this topic.

The lay stone was placed by Sir R. A. Fisher through his classic paper of 1918. Though he was not actually the first writer on this topic (Weinberg, 1909 and 1910), nevertheless Fisher's work remains a historical monument on its originality and comprehensiveness.

He used the principle of regression analysis to evaluate the correlations between relatives on the supposition of Mendelian inheritance. More or less in the same period Sewall Wright developed the concept of 'path coefficient' (standardized partial regression coefficient) which he thoroughly used to compute the various correlations of relatives (Wright 1918, 1920, 1921). It may be noted here that Fisher's concept of 'factors' of correlation is practically identical with that of 'path coefficients', though the two discoveries are independent (Li, 1968). Wright's method is basically a disguised form of the use of Bayes' rule and the law of total probabilities. Malécot (1948) recognized Wright's calculations by introducing the fundamental concept of identity by descent (for definitions see section 1.2) and exploiting its properties. The method of identity by descent has been perfected and developed by Malécot and his students, especially Gillois, Janquard and Bonffette. Li and Sacks (1954) fruitfully applied this concept to obtain the joint frequency distribution for any type of relatives in a simple manner. Kempthorne (1957) has applied the concept of identity by descent to the study of quantitative inheritance.

1.4 PARENT OFFSPRING CORRELATION IN GENERAL EQUILIBRIUM POPULATION

1.4.1 Two alleles at an autosomal locus without dominance relationship:

The correlation between parent and offspring, being the most important of all the genetic correlations, received the maximum attention in the literature. Fisher (1918) called such a correlation as 'parental correlation'. In a random mating population the joint distribution of parent-child pairs can easily be shown as designated in TABLE 1.4.1.

TABLE 1.4.1

Joint distribution of parent-offspring pair in a random mating population

			Child			
			AA	Aa	aa	Marginal total
		Z → ↓	2	1	0	
Parent	AA	2	p^3	p^2q	0	p^2
	Aa	1	p^2q	pq	q^2	$2pq$
	aa	0	0	pq^2	q^3	q^2
Marginal total			p^2	$2pq$	q^2	1

Note that this is the picture when we consider an autosomal character explained by two alleles at a locus, the three genotypes being AA, Aa and aa. For the time being we also assume that the action of the genes are additive (in the sense that there is no dominance relationship between the alleles). It is clear now that the variable Z relates to the gene-content of an individual (since it assigns the values 2, 1 and 0 to be the genotypes AA, Aa and aa respectively). Now it is clear enough that the correlation between the Z values of parent and offspring is (from TABLE 1.4.1)

$$r_{ZZ'} = \frac{1}{2}$$

which is independent of the allelic frequencies p and q in a population. The prime here refers to the child. The fact that $r_{ZZ'} = \frac{1}{2}$ is entirely a consequence of the Mendelian inheritance and the randomness of mating among the genotypes. This is the parent-offspring correlation. Note that this index is based upon pairs consisting of one parent and one child each. But when both parents and more than one offspring are measured, what sort of correlation should be adopted? This question was posed by Li (1954) and he showed that for random mating population the canonical correlation between the parental sets of variables and the offspring set of variables is the answer. In a general equilibrium population Li had no answer. Here study the nature of parent offspring correlations in a general equilibrium population as dictated by Model II.

Consider the case with two alleles at a locus where the frequencies of the different mating types in the population are given by TABLE 1.4.2 below. From the theorem 1.1.1 it follows that the population is in equilibrium if $u_{11} = 4\, u_{20}$. With Model II we have $D = p^2 + F\, pq$, $H = 2pq(1 - F)$ and $R = q^2 + Fq$ where p and q the frequencies of A- and a-alleles and F $(0 \leq F \leq 1)$ is a constant over generations measuring the degree of non-randomness.

TABLE 1.4.2

Frequency of matings in an equilibrium population

Mates	AA	Aa	aa	Total
AA	u_{22}	u_{21}	u_{20}	D
Aa	u_{21}	u_{11}	u_{10}	H
aa	u_{20}	u_{10}	u_{00}	R
Total	D	H	R	1

Following Li's argument (1954) here also the total scores of parents as well as offspring (s in number) are considered to derive the correlation.

The distribution of offspring - total for each mating can be obtained by considering the corresponding probability generating function (p.g.f.). For example let us consider the mating Aa x Aa. Here the p.g.f. of offspring total is $[f(x)]^s$ where p.g.f. of a single - offspring measurement is given by

$$f(x) = \frac{1}{4} + \frac{1}{2}x + \frac{1}{4}x^2$$

$$= \frac{1}{4}(1+x)^2, \quad |x| \leq 1.$$

Hence, Prob. (offspring - total = j | mating is Aa x Aa)

$$= \text{Co-eff. of } x^j \text{ in } \frac{1}{4^s}(1+x)^{2s}$$

$$= (\tfrac{1}{4})^s \cdot {}^{2s}C_j, \quad 0 \leq j \leq 2s.$$

Once the probabilities are obtained for all matings, the joint distribution of parental total and offspring - total can be obtained for any number of offsprings.

The joint distribution of these two total measurements when both parents and s offsprings are considered is specified by

$$P_{0,j} = \begin{cases} u_{00} & \text{if} \quad j = 0 \\ 0 & \text{if} \quad 1 \leq j \leq 2s \end{cases}$$

$$P_{1,j} = \begin{cases} 2u_{10} \cdot {}^{s}C_j \left(\frac{1}{2}\right)^s & \text{if} \quad 0 \leq j \leq s \\ 0 & \text{if} \quad s < j \leq 2s \end{cases}$$

$$P_{2,j} = \begin{cases} u_{11} \cdot {}^{2s}C_j \left(\frac{1}{4}\right)^s & \text{if} \quad 0 \leq j \leq s-1 \\ & \phantom{\text{if}} \quad s < j \leq 2s \\ 2u_{20} + u_{11} \cdot {}^{2s}C_s \left(\frac{1}{4}\right)^s & \text{if} \quad j = s \end{cases}$$

$$P_{3,2s-j} = \begin{cases} 2u_{21} \cdot {}^{s}C_j \left(\frac{1}{2}\right)^s & \text{if} \quad 0 \leq j \leq s \\ 0 & \text{if} \quad s < j \leq 2s \end{cases}$$

$$P_{4,j} = \begin{cases} u_{22} & \text{if} \quad j = 2s \\ 0 & \text{if} \quad 0 \leq j < 2s \end{cases}$$

where $P_{i,j}$ = Prob. (parental total = i, offspring - total = j).

The joint distribution of the offspring - total and parental - total when only one parent is available for measurement can be obtained from the above expressions identifying the reciprocal crosses and collecting cells according to the genotype of one parent (Note that for autosomal gene one-parent analysis is the same for mother or father).

From these joint distributions one can easily find out the variances and covariances of the two variables (parental total score and offspring total score) and hence the parent-offspring correlation. The details are omitted here and the expressions are presented in TABLE 1.4.3. Afterwards we shall present an alternative method of its computation which does not require these joint distributions. But these joint distributions dictate the relative frequencies of the families with all possible parental total and offspring total combinations.

For small s, the algebraic exercises can be carried out without the p.g.f. approach also. As s increases the correlation for one parent approaches $\sqrt{(1 + 3F) / [2(1 + F)]}$. A close observation enables us to see that for F = 0 (panmixia) this limiting value becomes $1/\sqrt{2}$ which is Li's observation. Thus, at this stage it is clear enough that Li's analysis (1954) is a special case of the present one.

To consider a special case, let us assume that w fraction of the population practice selfing in each generation and the rest (1 - w) fraction panmixia. It is known (Li, 1955) that it is equivalent to the

TABLE 1.4.3

Variance, covariance and correlations between parent(s) and child(ren)

	Number of children in a sibship = s
	$\sigma_c^2 = spq\,[(s+1) + (3s-1)F]$
One parent	$\sigma_{PC} = spq(1 + 3F)$
$\sigma_p^2 = 2pq(1 + F)$	$r = \dfrac{(1 + 3F)\sqrt{s}}{\sqrt{2(1+F)[(s+1) + (3s-1)F]}}$
Two parents	$\sigma_{PC} = 2spq(1 + 3F)$
$\sigma_p^2 = 4pq(1 + 3F)$	$r = \sqrt{\dfrac{s(1 + 3F)}{(s+1) + (3s - 1)}}$

σ_P^2 = Var. (Parent-total); σ_C^2 = Var. (Offspring-total)

σ_{PC} = Cov. (Parent-total, Off.— total) and r = corr. co. eff.

case with $F = \frac{w}{2 - w}$ .. Putting this in the asymptotic expression of correlation with single - parent measurement we get $r^{\cdot} = \sqrt{\frac{1 + w}{2}}$. This result is of particular interest for plant population. Correlation can be obtained for any number of offsprings inserting this value of F in the expressions of 1st row of TABLE 1.4.3.

Canonical Correlation :

The general treatment of this subject follows from the historic paper of Hotelling (1936) on canonical correlation. Denoting the measurements on parents by x_1 and x_2 and that for the j^{th} child (in order of birth) by y_j we have the expressions for variance and covariance as

$$\sigma^2_{x_1} = \sigma^2_{x_2} = \sigma^2_{y_1} = \ldots = \sigma^2_{y_s} = 2pq(1 + F) \qquad (1.4.1)$$

$$\sigma_{x_1 x_2} = 4\,F\,pq \qquad (1.4.2)$$

$$\sigma_{x_i y_j} = pq(1 + 3F) \qquad (1.4.3)$$

$$\sigma_{y_j y_{j'}} = pq(1 + 3F) \qquad (1.4.4)$$

Note that equations (1.4.1) and (1.4.2) are already in existence in the literature (Li, 1955 and Kempthorne, 1957). (1.4.3) is derived from the joint distribution of one offspring and one parental total and (1.4.4) is obtained as

$$\sigma_{y_j y_{j'}} = \frac{\sigma^2(y_j + y_{j'}) - 2\sigma^2_{y_j}}{2}$$

$$= \frac{2pq(3 + 5F) - 4pq(1 + F)}{2}$$

$$= pq(1 - 3F).$$

Now if $X = b_1 x_1 + b_2 x_2$ and $Y = c_1 y_1 + c_2 y_2 + \dots + c_s y_s$, the correlation between X and Y is given by

$$r_{XY} = \frac{(1 + 3F)\ (b_1 + b_2)(c_1 + c_2 + \dots + c_s)}{\sqrt{4[(b_1^2 + b_2^2)(1+F)+4Fb_1b_2][(c_1^2+\dots+c_s^2)(1+F)+(1+3F)\sum\limits_{j<j'} c_j c_{j'}]}}$$

which assumes a maximum value

$$r = \sqrt{\frac{s(1 + 3F)}{(s + 1) + (3s - 1)F}}$$

when all b's are equal and all c's are equal i.e.,

$$b_i = b \quad \text{and} \quad c_j = c \quad \text{for} \quad i = 1, 2$$
$$j = 1, 2, \dots, s.$$

The correlation between one parent and s children is

$$r = (1 + 3F)\sqrt{\frac{s}{2(1 + F)[(s + 1) + (3s - 1)F]}}$$

(putting either b_1 or b_2 equal to zero). Note that these are the same as those obtained in the last column of TABLE 2.

Direct Consequences :

Now let $r_{ij}(F)$ represent the canonical correlation coefficient when i parents and j children were considered ($i = 1, 2$; $j = 1, 2, \ldots, s, \ldots$). Then one can see that the following results are true :

(i) $r_{ij}(F) \geq f_{ik}(F)$ for $j > k$, $i = 1, 2$.

(ii) $r_{2j}(F) \geq r_{1j}(F)$ for all j

and (iii) $r_{ij}(F) \geq r_{ij}(F')$ for $F \geq F'$

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

These consequences prove that with inbreeding the parent-offspring corelation increases which is a natural finding.

Further putting $F = 1$ (complete fixation) in the general formula, all the correlations turn out to be unity which coincides with general theory.

A numerical illustration :

Taylor and Prior (1938) and Race et al. (1942) analysed two series of family data on MN blood groups from England. This illustration is also based on these two series of family data.

For M-N blood group system there are three phenotypically distinct genotypes MM, MN and NN. Assigning values 2, 1 and 0 (or equivalently M-gene content) to each individual, frequencies of parental total and offspring total are presented in the TABLE 1.4.4 for family sizes 1, 2, 3 and 4. Other families are left out from this analysis because of small numbers when classified according to the family size. The canonical correlations are also presented in the table. These provides a direct verification of the inequalities stated just now though family sizes 3 and 4 give a dismal result. This may as well be a case of sampling fluctuation. Of course, one may note that there are only 18 families with family size 4.

TABLE 1.4.4

Frequency distribution of the families according to the parental total and the offspring total
(s = 1)

Parental Total	Offspring Total			Total Frequency
	2	1	0	
4	5			5
3	6	15		21
2	3	15	2	20
1		8	13	21
0			1	1
Total Frequency	14	38	16	68

Correlation $(r_1) = 0.7077$

TABLE 1.4.4 (continued)

(s = 2)

Parental Total	Offspring Total 4	3	2	1	0	Total Frequency
4	7					7
3	5	11	9			25
2		4	12	5	1	22
1			4	11	2	17
0					4	4
Total Frequency	12	15	25	16	7	75

Correlation (r_2) = 0.8334

(s = 3)

Parental Total	Offspring Total 6	5	4	3	2	1	0	Total Frequency
4	3							3
3	2	3	2	1				8
2		1		16	2			19
1				5	6	3	1	10
0							4	4
Total Frequency	5	4	2	17	8	3	5	44

Correlation (r_3) = 0.9255

TABLE 1.4.4 (continued)

(s = 4)

Parental Total	Offspring Total									Total Frequency
	8	7	6	5	4	3	2	1	0	
4	1									1
3		1		2						3
2					5	2				7
1						3	1	1	1	6
0									1	1
Total Frequency	1	1		2	5	5	1	1	2	18

Correlation (r_4) = 0.9108

1.4.2 Two alleles at an autosomal locus with complete dominance :

In the section 1.4.1 we have assumed the action of the alleles A and a to be additive. Now let us relax that condition and assume A to be dominant over a so that only two phenotypes $\bar{A}$ (consisting of individuals with genotype AA and Aa) and $\bar{a}$ (consisting of individuals with genotype aa) are distinguishable. We assign a score of 1 and 0 to the individuals with phenotypes $\bar{A}$ and $\bar{a}$ respectively.

The phenotypic frequencies under Model II are given by

$1 - q^2 - Fpq$ and $q^2 + Fpq$ for $\bar{A}$ and $\bar{a}$ respectively} where p and q are A- and a-allele frequencies, $(p + q = 1)$, and F is the coefficient of non-randomness. For mathematical simplicity we assume that $p, q > F$ which is, of course, realized in most of the natural populations. With these, the three phenotypic mating types and their frequencies, as derived by Yasuda (1968), can be expressed as shown by TABLE 1.4.5. The u_{ij}'s, in the third column of this table, refer to the genotypic mating type frequencies as used by Li (1955) (e.g., u_{11} = Aa x Aa mating frequency; see also TABLE 1.4.2).

TABLE 1.4.5

Phenotypic mating types and their frequencies
(autosomal recessive character)

Mating type	Frequency	
	Yasuda' notation (1968)	Li's notation (1955)
$\bar{A} \times \bar{A}$	$p^2(1+q)^2 - 2Fpq(1-3q^2)$	$u_{22} + 2u_{21} + u_{11}$
$\bar{A} \times \bar{a}$	$2pq^2(1+q) + 2Fpq(1-6q^2)$	$2u_{20} + 2u_{10}$
$\bar{a} \times \bar{a}$	$q^4 + 6Fpq^3$	u_{00}

Once these are known, the joint distribution of the total scores of the parents and offspring (s in number) in a sibship is given by

$$\left.\begin{array}{l} P_{0,j} = \begin{cases} u_{00} & \text{if } j = 0 \\ 0 & \text{if } j > 0 \end{cases} \\ \\ P_{1,j} = \begin{cases} \binom{s}{j} \cdot (\frac{1}{2})^s \cdot 2u_{10} & \text{if } 0 \leq j < s \\ 2u_{20} + 2u_{10} \cdot (\frac{1}{2})^s & \text{if } j = s \end{cases} \\ \\ P_{2,j} = \begin{cases} u_{11} \cdot \binom{s}{j} \cdot (\frac{3}{4})^j \, (\frac{1}{4})^{s-j} & \text{if } 0 \leq j < s \\ u_{22} + 2u_{21} + (\frac{3}{4})^s \, u_{11} & \text{if } j = s \end{cases} \end{array}\right\} \tag{1.4.5}$$

where, $P_{i,j}$ = Prob. (parental - total = i, offspring - total = j) for $i = 0, 1;\ j = 0, 1, 2, \ldots, s$.

Denoting the scores for parents by x_1 and x_2 and that for the j^{th} offspring (by the order of birth) by y_j we have the variances and covariances given by the expressions as follows:

$$\sigma^2_{x_1} = \sigma^2_{x_2} = \sigma^2_{y_1} = \ldots\ldots = \sigma^2_{y_s} = A.pq \tag{1.4.6}$$

$$\sigma_{x_1x_2} = (A - B)pq \tag{1.4.7}$$

$$\sigma_{x_iy_j} = [A - (q + F - 3Fq)]pq \tag{1.4.8}$$

and

$$\sigma_{y_jy_{j'}} = \frac{pq}{4}[4A - C] \tag{1.4.9}$$

where, $A = (q + Fp)(1 + q - Fq)$

$B = F + q + q^2 - 6Fq^2$

and $C = 3F + 4q - pq - 6Fq - 6Fq^2$.

Derivation of (1.4.6) is obvious since each of the x and y variables takes only two values 1 and 0 with probabilities $(D + H)$ and R respectively and hence variance $= R(D + H) = A.pq$. The covariances given by equations (1.4.7) to (1.4.9) can easily be worked out by considering the suitable joint distributions.

Using these equations (or, directly from (1.4.5)) we obtain the correlation between parental total score, $X \left(= x_1 + x_2\right)$ and offspring total score, $Y \left(= \sum_{j=1}^{s} y_j\right)$ as

$$r_{XY} = \frac{\sqrt{2s} \, . \, q \left[q(1 - F)^2 + F(3 - F) \right]}{\sqrt{\left[(2A - B)\left[s.A - \frac{C(s - 1)}{4} \right] \right]}} \qquad (1.4.10)$$

Putting $F = 0$ in (1.4.10), we get the parent-offspring correlation (r^*_{XY}) for random mating population as

$$r^*_{XY} = \frac{2 \sqrt{2} \, . \, pq}{\sqrt{\left[(1 + q)\left[(3 + q) + s(4 + pq) \right] \right]}} \qquad (1.4.11)$$

In this case also it can be shown that the expression in (1.4.10) represents the maximum correlation between the two sets of scores (i.e., the set of parental scores and the set of offspring scores). The treatment is analogous to the undominated case, as described earlier.

A numerical illustration :

For this purpose, we consider the data of Race et al. (1948, 1949) where they subdivided the MN blood group system by using iso-agglutinin S. Though the discovery of s-serum now divides the individuals into three phenotypically distinct genotypes SS, Ss and ss, we consider the grouping with only S-serum and thus get phenotypes S+ (consisting of SS and Ss individuals) and S- (consisting of ss individuals). The joint distributions of parental total score and offspring total scores are shown by Table 1.4.5 for s = 1, 2, 3 and 4 separately. The parent offspring correlation is also presented below the tables for different family sizes separately. Expression (1.4.10) also yields the same correlations with a F-value of 0.05 since the frequencies of S and s alleles are 0.3812 and 0.6177 respectively.

TABLE 1.4.5

Frequency distribution of families according to parental total and offspring total scores

(s = 1)

Parental Total	Offspring total		Total Frequency
	0	1	
0	8	0	8
1	6	11	17
2	1	6	7
Total Frequency	15	17	32

Correlation $(r_1) = 0.5809$

TABLE 1.4.5 (contd.)

(s = 2)

Parental Total	Offspring total			Total Frequency
	0	1	2	
0	5	0	0	5
1	2	5	6	13
2	0	5	17	22
Total Frequency	7	10	23	40

Correlation (r_2) = 0.7036

(s = 3)

Parental Total	Offspring total				Total Frequency
	0	1	2	3	
0	4	0	0	0	4
1	0	8	9	4	21
2	0	0	2	7	9
Total Frequency	4	8	11	11	34

Correlation (r_3) = 0.7676

TABLE 1.4.5 (contd.)

(s = 4)

Parental Total	Offspring total					Total Frequency
	0	1	2	3	4	
0	3	0	0	0	0	3
1	0	1	1	0	1	3
2	0	0	0	3	1	4
Total Frequency	3	1	1	3	2	10

Correlation $(r_4) = 0.8548$

1.5 PARENT OFFSPRING CORRELATIONS FOR SEX LINKED CHARACTERS

In section 1.4 we have derived the canonical correlation between the parental set of scores and offspring set of scores when the character concerned is explained by two alleles at an autosomal locus. The sex-linked characters need special attention since the asymmetrical chromosomal complement of males and females makes it necessary to distinguish the sexes of the individuals. We shall take as usual the homogametic type XX as females and the heterogametic XY (or XO) as males, where X denotes the sex chromosome. Hence, with two alleles A and a resting on a locus on this sex chromosome we have phenotypes A and a for males and AA, Aa and aa for females when the alleles are codominant so far as their gene actions are concerned. In such a situation the necessary and sufficient condition for general genetic equilibrium is given by Theorem 1.1.2. It is easy to see that under Model II the genotypic frequencies are given by $(p.A + q.a)$ ♂ and $\left[(p^2 + Fpq)AA + (2pq - 2Fpq)Aa + (q^2 + Fpq)aa\right]$ ♀ where, p and q are the A- and a-allele frequencies respectively. F has the same interpretation as in the other cases. The six different mating types, their frequencies and the segregation ratios can be represented by TABLE 1.5.1. Note that Theorem 1.1.2 dictates the equilibrium condition given by $u_{11} = 2u_{20}$ and $u_{10} = 2u_{01}$ which is satisfied by the mating type frequencies under Model II.

TABLE 1.5.1

Mating types, their frequencies and the segregation ratios (sex - linked codominant character)

Genotypic Mating type	Frequency		Segregation ratio				
	Yasuda (1968)	Chakraborty (1970)	Boys		Girls		
♀ ♂			A	a	AA	Aa	aa
AA x A	$p^3 + 3Fp^2q$	u_{21}	1	0	1	0	0
AA x a	$p^2q + Fpq(1 - 3p)$	u_{20}	1	0	0	1	0
Aa x A	$2p^2q + 2Fpq(1 - 3p)$	u_{11}	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0
Aa x a	$2pq^2 + 2Fpq(1 - 3q)$	u_{10}	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$
aa x A	$pq^2 + Fpq(1 - 3q)$	u_{01}	0	1	0	1	0
aa x a	$q^3 + 3Fpq^2$	u_{00}	0	1	0	0	1

As in the earlier case, we denote here the scores on mother and father by x_1 and x_2 respectively and that on j^{th} son and k^{th} daughter (in the order of birth) by y_j and z_k respectively. Then we have the variances and covariances between them as

$$\left.\begin{aligned}
&\sigma^2_{x_1} = \sigma^2_{z_1} = \sigma^2_{z_2} = \ldots = \sigma^2_{z_r} = 2(1 + F)pq\\
&\sigma^2_{x_2} = \sigma^2_{y_1} = \sigma^2_{y_2} = \ldots = \sigma^2_{y_s} = pq\\
&\sigma_{x_2 y_j} = 0\\
&\sigma_{x_1 y_j} = \sigma_{x_2 z_k} = (1 + F).pq\\
&\sigma_{y_j y_{j'}} = \tfrac{1}{2}(1 + F)pq\\
&\sigma_{z_k z_{k'}} = \tfrac{1}{2}(3 + 5F).pq\\
&\sigma_{x_1 x_2} = 2Fpq\\
&\sigma_{x_1 z_k} = (1 + 3F)pq\\
&\sigma_{y_j z_k} = \tfrac{1}{2}(1 + 3F)pq
\end{aligned}\right\} \qquad (1.5.1)$$

The derivations of these expressions are analogous to those of section 1.4 and hence are omitted. However, these are obvious once the suitable joint distributions are recorded. The joint distributions can easily be constructed from the TABLE 1.5.1.

If, now, we denote the parental total score by X $(= x_1 + x_2)$, total score of s number of sons by Y $(= \sum_{1}^{s} y_j)$ and the total score of r number of daughters by Z $(= \sum_{1}^{s} z_k)$, we have correlation between X and Y as

$$r_{XY} = \left[\frac{2s(1 + 3F)}{3[(s + 1) + (s - 1)]} \right]^{\frac{1}{2}} \tag{1.5.2}$$

and correlation between X and Z as

$$r_{XZ} = 2 . \left[\frac{3r(1 + 2F)}{3[(3r + 1) + (5r - 1)F]} \right]^{\frac{1}{2}} \tag{1.5.3}$$

When only one parent is measured, the different correlations are seen to be as follows :

$$\begin{aligned} r_{x_1 Y} &= \left[\frac{s(1 + 3F)}{(s + 1) + (s - 1)F} \right]^{\frac{1}{2}} \\ r_{x_1 Z} &= \left[\frac{r(1 + 3F)^2}{(1 + F)[(3s + 1) + (5s - 1)F]} \right]^{\frac{1}{2}} \end{aligned} \tag{1.5.4}$$

$$r_{x_2 Y} = 0$$

and $$r_{x_2 Z} = (1 + F) \left[\frac{2r}{(3r + 1) + (5s - 1)F} \right]^{\frac{1}{2}}$$

In a family with s number of sons and r number of daughters, the parent offspring correlation $[r_{X,(Y + Z)}]$ is seen to be

$$r_{X,Y+Z} = [s(1 + F) + 2r(1 - 2F)] . \left[\frac{2}{3D(1 + 2F)} \right]^{\frac{1}{2}} \tag{1.5.5}$$

where, $D = 2rs(1 + 3F) + r[(3r + 1) + (5r - 1)F] + s[(s + 1) + (s - 1)F]$.

When the character is sex-linked recessive in nature the variances and covariances of the x, y and z variables are given by the expressions in Table 1.5.2. The table also gives an illustration by using the data of Adam et al (1963 and 1967) on Xg blood group. It is well known that antigen Xg^a behaves as an X-linked dominant one and thus their tests were not efficient at distinguishing heterozygous Xg^aXg, females from those homozygous, Xg^aXg^a. The observed variances and covariances are quite in good agreement with the expected ones (obtained by using the gene frequency estimates of Adam et al, 1967; $p = 0.678$ and $q = 0.322$ and a F-value of the magnitude of 0.02).

From the expressions of TABLE 1.5.2 one can, easily, obtain all the parent offspring correlations. Needless to say that the corresponding expressions for panmictic populations can be obtained by putting $F = 0$. Thus we can have the expressions for different parent offspring correlations for Model II as well as Model I.

It may be recalled here that for autosomal genes the correlations between four types of parent-child pairs (consisting of one parent and one child each) are the same, all being $(1 + 3F)/[2(1 + F)]$. But for sex-linked genes, because of the reason already mentioned, they are not the same. We have already noticed that there are four kinds of parent-offspring relationships: father-son, father-daughter, mother-son and mother-daughter. Such correlations will be studied in greater details in section 1.7.

TABLE 1.5.2

Variances and covariances with sex-linked recessive model

Quantity		Value for Xg blood group data	
		Expected*	Observed
Variance : father	pq	0.2183	0.2208
Variance : son	pq	0.2183	0.2169
Variance : mother	$A.pq$	0.1256	0.1261
Variance : daughter	$A.pq$	0.1256	0.1255
Father - son covariance	0	0.0	-0.0042
Mother - son covariance	$pq(q + Fp)$	0.0999	0.0895
Father - daughter covariance	$pq(q + Fp)$	0.0999	0.0996
Father - mother covariance	$2Fpq$	0.0873	0.0886
Mother - daughter covariance	$A.pq-pq(F+q-3Fq)$	0.0931	0.0896
Brother - sister covariance	$\frac{pq}{2}[q + (1 + q)F]$	0.0640	0.0653

* Expected values are obtained by putting $p = 0.0678$, $q = 0.322$ (Adam et al, 1967) and $F = 0.02$

1.6 PARENT OFFSPRING CORRELATIONS FOR MULTI-ALLELIC CASE

In the two allelic case we have assigned weights 0, 1 and 2 to the genotypes aa, Aa and AA respectively and thence computed the correlations. This arbitrary assignment of weights is a natural formulation of the fact that the heterozygote Aa is intermediate between the homozygotes. But this superficial approach leads us to encounter difficulty when we attempt to extend the procedure to n (≥ 3) alleles. That is how the multi-allelic case needs special attention. Herein first we discuss two weighting schemes developed by Stanton (1960) and using these we indicate how the foregoing analysis can be readily extended for multiple alleles.

Genetic Weighting :

Suppose that we consider the case of autosomal genes wherein at a locus there rest n alleles $A_1, A_2, \ldots, A_n$. For understanding the need for a genetic weighting scheme let us analyse the case with $n = 3$. A weighting method like that of 2-allelic case leads us to arrange the genotypes as A_1A_1, A_1A_2, A_2A_2, A_2A_3, A_3A_3, A_3A_3, and attach weights 0, 1, 2, 3, 4 and 5. Thus the genotype A_3A_1 receives the weight of magnitude 5 which is unfortunately not intermediate between the two homozygotes A_1A_1 and A_3A_3 having weights 0 and 4 respectively. This anomaly is not surprising enough if only one notes that the above weights assigned to the genotypes do not represent genetic effects per se. Therefore we feel the need of a weighting scheme whereby the homozygotes are placed symmetrically with respect to one another. The above weighting design places the homozygotes linearly on the weighting axis only and does not possess the required symmetry.

These considerations suggest that one might use the complex vectorial weights and in some way bring the symmetry of the site of the homozygotes. Suppose the n homozygotes are placed at the vertices of a regular simplex with n vertices in $(n - 1)$-dimensional space; it is then natural to place the heterozygotes at the mid-points of the edges. For symmetry, we may suppose that the origin of coordinates is at the centroid of the simplex. Let v_i denote the vector joining the origin to the point representing A_iA_i. We assign this vectorial weight to the genotype A_iA_i. To the heterozygote A_iA_j we assign a weight of $\frac{1}{2}(v_i + v_j)$ which represents the vector joining the origin to the point where A_iA_j is located. With these vectorial weights an algebra is generated where the vector products are interpreted as scaler products and with this convention Stanton (1960) had

$$v_i^2 = v_i \cdot v_i = \text{Constant} . \ (i = 1, 2, \ldots, n) \qquad (1.6.1)$$

For simplicity, this constant is taken as 1 (i.e., the vectors joining the origin to the n vertices are each of unit length). Then, using the fact that all of the edges of a regular simplex are of equal length, we have

$$v_i\, v_j \;=\; \text{constant } (i \neq j). \tag{1.6.2}$$

Also we have

$$v_1 + v_2 + v_3 + \ldots + v_n = 0 \tag{1.6.3}$$

since the centroid of the simplex is taken as the origin.

Multiplying (1.6.3) by v_i we have $v_i^{\,2} + (n - 1)\, v_i v_j = 0$ and hence using (1.6.1) we have

$$v_i\, v_j = -1/(n - 1) \quad (i \neq j). \tag{1.6.4}$$

The idea can be illustrated, for three allelic case, by Figure 1.6.1 wherein the simplex has a physical realization in the sense that it is an equilateral triangle and the weights are ordinary complex numbers. Using expressions (1.6.1) and (1.6.4) we have for this case

$$v_i^{\,2} = 1,\; v_i v_j = -\tfrac{1}{2},\; \tfrac{1}{2}\, v_i(v_j + v_k) = -\tfrac{1}{2}$$

$$\tfrac{1}{2}\, v_i(v_i + v_j) = \tfrac{1}{4},\; \left[\tfrac{1}{2}(v_i + v_j)\right]^2 = \tfrac{1}{4},\; \tfrac{1}{2}(v_i + v_j)\,\tfrac{1}{2}(v_i + v_k) = \frac{1}{8} \quad (1.6.5)$$

Throughout the relations (1.6.5), i, j and k range from 1 to 3

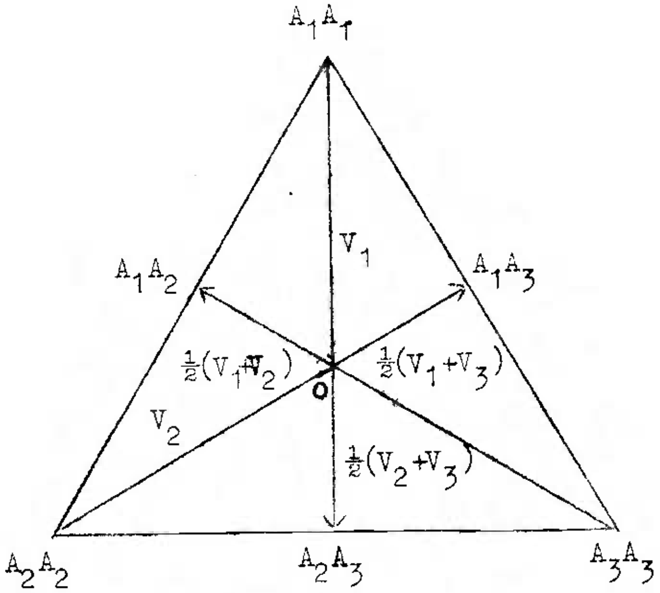

FIGURE 1.6.1

Weighting scheme for three allelic case (after Stanton, 1960)

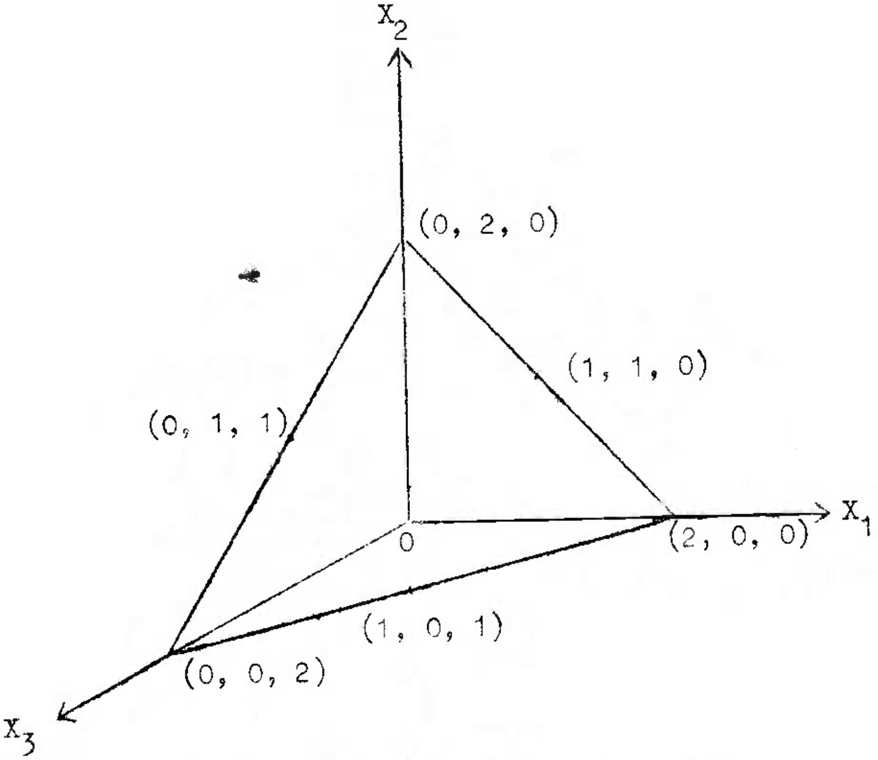

FIGURE 1.6.2

Alternative weighting scheme for three alleles (after Stanton, 1960) but are distinct from one another in any single relation.

An alternative weighting scheme : Stanton (1960) developed yet another approach to solve the weighting problem wherein he characterized the genotypes with n alleles at a locus by vector variables $(X_1, X_2, \ldots, X_n)$ where each X_i is 0, 1 or 2, and thus denotes the number of A_i genes in the genotype. Thus for each such vector variable we have $X_1 + X_2 + \ldots + X_n = 2$. Thus the genotypes with such weights attached to them will all lie on the hyperplane $X_1 + X_2 + \ldots + X_n = 2$. Such a situation in the 3-allelic case can be illustrated with Figure 1.6.2. The difference between Figure 1.6.1 and 1.6.2 is that the origin in Figure 1.6.2 has been projected into the centroid of the triangle and a change of scale has been introduced in order to normalize distances and make the distance from the centroid to each vertex equal to unity.

It may be mentioned here that this alternative weighting scheme is not used to derive the correlations. This is mentioned only for the sake of completeness.

Parent-Offspring Correlation :

We illustrate the method of extension taking $n = 3$ for ease in illustration. The results are general, although the proof for arbitrary n is simpler by a stochastic method. Let the frequency of gene A_i be denoted by q_i, with the customary convention that the total frequency be unity, that is, $q_1 + q_2 + q_3 = 1$.

The frequencies of the different genotypes in the case of autosomal alleles without any dominance relationship between them are given by Model II (expression (1.1.2)) and the weights attached to them having relationships expressed by (1.6.5). Then the average weight attached to individuals of the population is given by

$$\begin{aligned}&\sum_i v_i \left[q_i^2 + Fq_i(1 - q_i) \right] + 2(1 - F) \sum_{i<j} q_i q_j \cdot \tfrac{1}{2}(v_i + v_j) \\ &\quad = \sum_i q_i v_i \end{aligned} \qquad (1.6.6)$$

and hence the variance of the weights is given by

$$\begin{aligned}&\sum_i v_i^2 \left[q_i^2 + Fq_i(1 - q_i) \right] + 2(1-F) \sum_{i<j} q_i q_j \left[\tfrac{1}{2}(v_j + v_j) \right]^2 - \Big(\sum_i q_i v_i \Big)^2 \\ &\quad = \frac{3}{2} (1 + F) \sum_{i<j} q_i q_j \,. \end{aligned}$$

Thus, with the same interpretations of x's and y's as before, one now obtains

$$\left. \begin{aligned} &\sigma_{x_1}^2 = \sigma_{x_2}^2 = \sigma_{y_1}^2 = \ldots = \sigma_{y_s}^2 = \frac{3}{2}(1 + F) \sum_{i<j} q_i q_j \\ &\sigma_{x_i y_j} = \sigma_{y_i y_j} = \frac{3}{4}(1 + F) \sum_{i<j} q_i q_j \\ \text{and} \quad &\sigma_{x_1 x_2} = 3F \sum_{i<j} q_i q_j \end{aligned} \right\} \qquad (1.6.7)$$

which leads to the correlation between X and Y as

$$r_{XY} = \left[\frac{s(1 + 3F)}{(1 + s) + (3s - 1)F} \right]^{\frac{1}{2}}$$

as found in the two allelic case.

In case the character is sex-linked, the males can have only 3 possible genotypes, that is, A_1, A_2 or A_3. We immediately find that, for females, the mean and variance are the same as before, as given by (1.6.6) and (1.6.7). For males, we find that

$$\text{mean} = \sum_i q_i v_i$$

$$\text{and} \quad \text{variance} = \sigma_{x_2}^2 = 3 \sum_{i<j} q_i q_j \qquad (1.6.8)$$

One can now obtain the covariances easily and hence get the parent offspring correlation. It is interesting to note that like the autosomal case, here also the expression are in complete agreement with the two allelic case.

The case with dominance is studied in some details by Stanton (1960) though his results are only applicable to panmictic populations described by Model I. He referred to three possible types of dominance as follows :

Type I : A_1 dominates A_2 and A_3, A_2 dominates A_3.

Type II : A_1 dominates both A_2 and A_3, neither A_2 nor A_3 dominates.

Type III : A_1 and A_2 dominate A_3, neither A_1 nor A_2 dominates the other.

In the case with dominance, even under panmixia, one cannot employ the method of stochastic matrices because of some inherent difficulties. However, by the direct approach the corresponding expressions can easily be worked out analogously.

1.7 CORRELATIONS BETWEEN OTHER RELATIVES IN AN EQUILIBRIUM POPULATION

Correlations between the other relatives are also as important as parent-offspring correlation so far as the use of these correlations to study the mechanism of inheritance is concerned. And very often the knowledge of absolute frequencies of the various genotypic combinations of any two relatives with respect to one pair of genes, autosomal or sex-linked, is regarded as more important in certain types of studies in human heredity. Various methods of obtaining these frequencies directly or indirectly are prevailing in the literature. The difficulties of the rather straight forward procedure (that is, obtaining the joint distribution from the mating table) can be well apprehended from Hogben's paper (1933). Li and Sacks (1954) gave a procedure of finding the frequencies of various genotype combinations of near relatives by using the matrices of conditional probabilities. The main purpose of their method, well known as ITO method, was to express such matrices of conditional probabilities of the relatives in the form of linear combination of some basic matrices. A more precise form of their work can be seen in Karlin's review paper (1968) where he has emphasized that this method is really an application of the concept of identity by descent as developed by Malecot (1948).

In case one is more interested in the correlations between the relatives rather than the frequencies of different genotype/phenotype combinations, the method of path coefficients, developed by Wright (1921 and later) pays dividend since it gives the correlations instantly once the relationship is specified. This method, of course, does not provide the absolute frequencies of the different genotype (or phenotype) combinations.

But all these methods are well known to study the genetic correlations between relatives when the population is assumed to exercise random mating only. Relatively fewer results are known for populations which keep themselves under equilibrium through a more general mating structure. In this section we derive the genetic correlations between some near relatives in a population described by Model II. We shall study the different cases with one pair of alleles at a locus in detail and indicate the necessary generalizations in the case with multiple alleles at a locus.

Two codominant alleles at an autosomal locus : Let A and a denote the two codominant alleles at an autosomal locus. Now, recalling the definitions 1.2.1 and 1.2.2, we note that a pair of relatives may have both, one or no gene identical by descent depending upon the type of relationship. In case there is exactly one gene common through identity by descent the conditional probabilities that one should be of a certain genotype when the other's genotype is given can be represented by the matrix

$$T_* = \begin{bmatrix} p\left[1 + \dfrac{2Fq}{p + Fq}\right] & q\left[1 - \dfrac{2Fp}{p + Fq}\right] & 0 \\ \frac{1}{2}\left[p + \dfrac{F(1 - 2p)}{1 - F}\right] & \frac{1}{2} & \frac{1}{2}\left[q + \dfrac{F(1 - 2q)}{1 - F}\right] \\ 0 & p\left[1 - \dfrac{2Fq}{q + Fp}\right] & q\left[1 + \dfrac{2Fp}{q + Fp}\right] \end{bmatrix}$$

where p and q are the A- and a-allele frequencies and F has the same interpretation as in the earlier case. One may observe that this also represents the matrix of transition probabilities of a parent offspring pair since a parent and one of its offspring always share one gene through identity by descent.

Hence, multiplying the first row of T_* by $p^2 + Fpq$, second row by $2pq(1 - F)$ and the third row by $q^2 + Fpq$ one can convert T_* into absolute frequencies for different genotypic combinations of a parent offspring pair from which the parent offspring correlations (r_{T_*}) is obtained as

$$r_{T_*} = \frac{1 + 3F}{2(1 + F)} \tag{1.7.1}$$

Since we are, for the moment, concerned with autosomal genes only, this will be the correlation for all parent offspring combinations (mother-son, mother-daughter, father-son or father-daughter). With this correlation alone we can obtain the correlation between the full sibs or two half sibs through Figure 1.7.1 and 1.7.2 as follows :

The correlation between two full sibs

$$r_{FS_*} = 2\ r_{T_*}^2\ (1 + m) \tag{1.7.2}$$

where m, the correlation between the mating partners is given by $m = 2F/(1 + F)$.

Inserting this value of m in the above expression (1.7.2) one gets

$$r_{FS_*} = \tfrac{1}{2}\left[\frac{1 + 3F}{1 + F}\right]^3 \qquad (1.7.3)$$

Putting $F = 0$ in the expression of (1.7.3) we obtain the correlation between the full sibs in a random mating population as

$$r_{FS} = 2.(\tfrac{1}{2})^2 = \tfrac{1}{2} .$$

In Figure 1.7.2 we denote λ to be the correlation between the parents who mate with the common parent. (In the figure, λ is the correlation between the two wives; the busband being the common parent of the two half sibs).

The correlation between the half sibs turns out to be

$$r_{HS_*} = r_{T_*}^2 \cdot \left[(1 + m)^2 + \lambda\ (1 + 2m)\right]. \qquad (1.7.4)$$

Correlations between other relatives can also be worked out similarly using this basis correlation r_{T_*}.

Genes with dominant relationship at an autosomal logus :

Let us now assume that the allele A to be dominant over a. Thus we have with us phenotypes $\bar{A}$ and $\bar{a}$. The parent-offspring correlation takes the form :

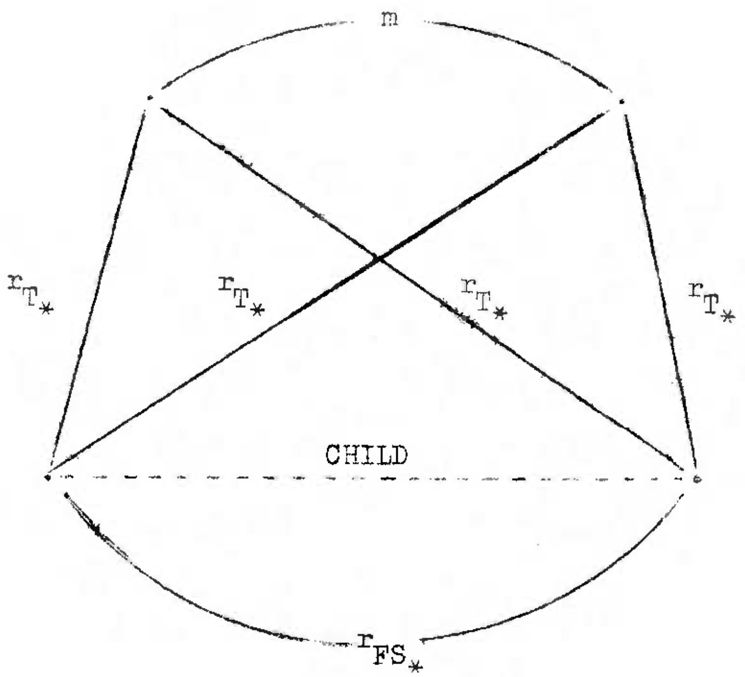

FIGURE 1.7.1

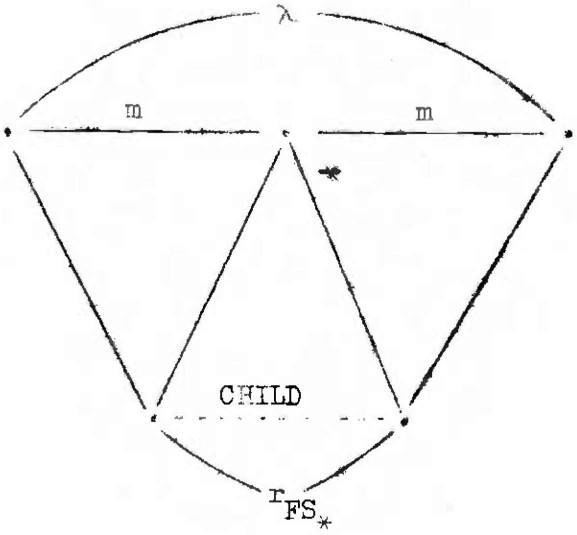

FIGURE 1.7.2

$$r_{T_*}^{(d)} = \frac{q}{1 + q - Fq} + \frac{Fq(1 + p - Fp)}{(1 + q - Fq)(q + Fp)} \qquad (1.7.5)$$

Needless to say that for $F = 0$ (Model I) this expression reduces to $q/(1 + q)$ as obtained by Li (1955).

The full sib correlation is easily seen to be

$$r_{FS_*}^{(d)} = \frac{1}{4} + \frac{1}{2}\, r_{T_*}^{(d)} = \frac{1}{4} + \frac{1}{2}\left[\frac{q}{1 + q - Fq} + \frac{Fq(1 + p - Fp)}{(1 + q - Fq)(q + Fp)}\right] \qquad (1.7.6)$$

It may be recalled that in the absence of any dominance relationship between the alleles, the genetic correlations depend only on F, the coefficient of departure from random mating. The correlations are independent of gene frequencies. But in this case the correlations do depend upon the gene frequencies as well. Under panmixia the correlation between the full sibs is given by

$$r_{FS}^{(d)} = \frac{1}{4} + \frac{1}{2}\left(\frac{q}{1 + q}\right).$$ [putting $F = 0$ in the expression (1.7.6)].

Codominant alleles at a sex-linked locus*:

We have already seen that due to the asymmetric chromosomal complement of males and females it is necessary to distinguish the sexes of the relatives in such a case and hence we need to consider all of the four kinds of parent offspring relationships : father-son, father-daughter, mother-son and mother-daughter. Likewise there are three kinds of sib-pairs : two brothers, two sisters and brother-sister.

The parent offspring correlations can easily be obtained once we consider the segregation ratios as shown in TABLE 1.5.1. The corrëlations are as follows :

$$\left.\begin{aligned}
&\text{Father - son correlation, } r_{fs} = 0.\\
&\text{Father - daughter correlation, } r_{fd} = \left(\frac{1+F}{2}\right)^{\frac{1}{2}}\\
&\text{Mother - son correlation, } r_{ms} = \left(\frac{1+F}{2}\right)^{\frac{1}{2}}\\
\text{and}\quad &\text{Mother - daughter correlation, } r_{md} = \frac{1+3F}{2(1+F)}
\end{aligned}\right\} \qquad (1.7.7)$$

From (1.7.7) we get interestingly enough that there is no correlation between father and son for sex-linked genes, since the son receives his father's Y-chromosome only which is void of the locus under consideration. The correlation for mother-daughter is the same as in the case of autosomal genes, because the daughter also receives p.A and q.a from her father. The two correlations for father-daughter and mother-son are the same. Note that each of these statements are true also in case of random mating populations.

Once these basic correlations are obtained the full sib correlations are obtained from the Figure 1.7.3 as follows :

$$\left.\begin{aligned}
&\text{Brother - brother correlation, } R_1 = \frac{1 + F}{2} \\
&\text{Brother - sister correlation, } R_2 = \frac{1 + 5F + 2F^2}{2\sqrt{2(1 + F)}} \\
&\text{Sister - sister correlation, } R_3 = \frac{1 + F}{2} + \frac{(1 - 3F)(1 + 7F + 4F^2)}{4(1 + F)^2}
\end{aligned}\right\} \quad (1.7.8)$$

In deriving this we made use of the fact that correlation between the mating partners is given by

$$m = F\sqrt{2} \,/\, \sqrt{1 + F}$$

in this case.

From Figures 1.7.4 and 1.7.5 one can now establish the following relations for half-sib correlations :

$$\left.\begin{aligned}
&\text{Brother - brother (related through mother) correlations,} \\
&\qquad R_4 = \frac{1 + F}{2} \\
&\text{Brother - sister (related through mother) correlations,} \\
&\qquad R_5 = \frac{1 + F}{2}\left[\lambda_1 F + \frac{1 + 5F + 2F^2}{2(1 + F)}\right] \\
&\text{Sister - sister (related through mother) correlation,} \\
&\qquad R_6 = \left[\frac{1 + 5F + 2F^2}{2(1 + F)}\right]^2 + \frac{\lambda_1(1 + 4F + 7F^2)}{2(1 + F)}
\end{aligned}\right\} \quad (1.7.9)$$

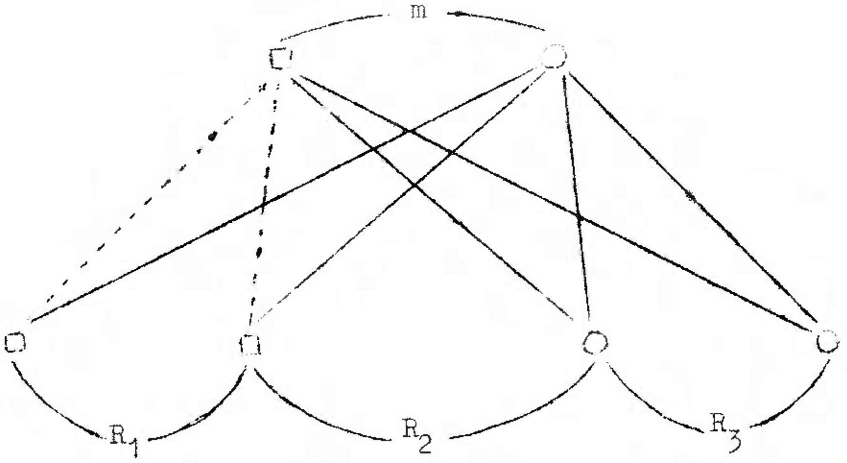

FIGURE 1.7.3

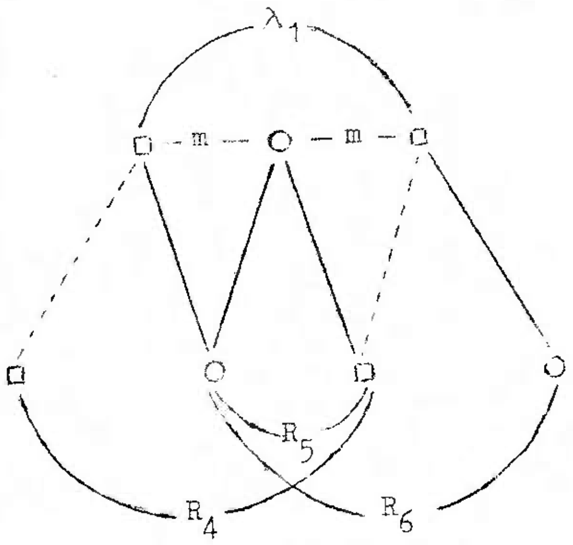

FIGURE 1.7.4

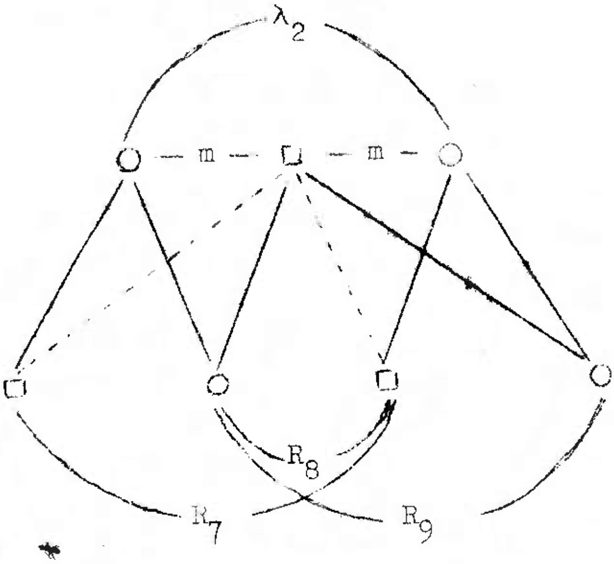

FIGURE 1.7.5

Brother - brother (related through father) correlation,

$$R_7 = F^2 + \frac{\lambda_2(1 + F)}{2}$$

Brother - sister (related through father) correlation,

$$R_8 = \sqrt{\frac{1 + F}{2}} \left[F(1 + \lambda_2) + \frac{1 + 3F}{1 + F} \left(\frac{F^2}{1 + F} + \frac{\lambda_2}{2} \right) \right] \tag{1.7.10}$$

and sister - sister (related through father) correlation,

$$R_9 = \frac{1 + F}{2} + \frac{F(1 + \lambda_2)(1 + 3F)}{1 + F} + \frac{(\lambda_2 + \lambda_2 F + 2F^2)(1 + 3F)^2}{4(1 + F)^3}$$

where λ_1 and λ_2 are the correlations between the two fathers (figure 1.7.4) and two mothers (figure 1.7.5) respectively.

When the allele A is dominant over a at a sex-linked locus, by similar argument one can show that the basic parent-offspring correlations

$$r_{md} = 1 - \frac{(F - q - 3Fq)}{(1 + q - Fq)(q + Fp)}$$

$$r_{ms} = r_{fs} = \sqrt{\frac{q + Fp}{1 + q - Fq}} \tag{1.7.11}$$

and $r_{fs} = 0.$

Note that the facts stated in the case of autosomal genes with dominant relationship also holds good in this case.

Multiple alleles at a segregating locus :

The foregoing analysis can readily be extended to multiple alleles, autosomal or sex-linked. But the superficial weighting system is not enough for it because of the reasons mentioned earlier. Using Stanton's weighting scheme it can be seen that the parent-offspring correlations, and hence the other correlations, turn out to be the same as those in the two allelic case. Worth noting that Stanton (1960) also observed this for a Model I population.

We now present an illustration of parent-offspring correlation for characters controlled by one pair of genes with dominance and an hypothetical case with a three allelic autosomal character. For both of them Boorman's data (1950) on human blood factors are sufficient.

Though the genetics of the Rh-factor in human blood has advanced greatly in the last two decades for our purpose here we may, however, still treat it as though it were controlled by one pair of genes with dominance. Boorman (1950) reported the data regarding mother-child combinations as shown in TABLE 1.7.1 here.

The parent-offspring correlation (in this case mother-child correlation) is 0.2853. The theoretical formula (1.7.5) also leads to the same amount of correlation with a F value of 1.8 percent.

TABLE 1.7.1

Mother - child combination of Rh blood factors

(After Li, 1955)

Mothers	Rh type of child		Total
	(+)	(-)	
(t)	1475	182	1657
(-)	204	129	333
Total	1679	311	1990

Mother - child correlation = 0.2853

To illustrate the three allelic case we consider TABLE 1.7.2 which is constructed from Boorman's data (1950). Using the same example Stanton (1960) computes a correlation of magnitude 0.5180 which leads to an estimate of F = 1.86 percent.

TABLE 1.7.2

Parent - child array for A-B-O blood types

(After Stanton, 1960)

	Child						
Parent	AA	AB	BB	BO	OO	AO	Total
AA	49	5	-	-	-	122	176
AB	10	10	2	21	-	24	67
BB	-	22	-	4	-	-	6
BO	-	20	4	50	56	29	159
OO	-	-	-	43	622	227	892
AO	122	14	-	28	223	303	690
Total	181	151	6	146	901	705	1990

1.8 PARENT OFFSPRING CORRELATION AND ESTIMATION OF F

Starting from Sewall Wright, to whom the concept of F is due, there had been many contributions towards the direction of estimating this important parameter. Estimation of F is of obvious interest to geneticists who wish to understand population structure and evolutionary processes which determine the associated array of genotypes. Moreover since this parameter is commonly interpreted as the coefficient of inbreeding considerable effort has been directed toward delineating the effect of inbreeding on quantitative as well as qualitative characters. And no such attempt to appreciate these effects could be fruitful without mentioning of the estimate of F, the coefficient of inbreeding. But it should be noted that the total amount of inbreeding in a population F_T, can be partitioned into a portion, F_A, that can be ascertained by study of records indicating consanguinity existing within a population, and a portion, F_R or remote consanguinity, which is undetectable by an analysis of pedigrees or other records, such that

$$F_T = F_A + F_R \qquad \text{(Morton and Yasuda, 1962)}$$

Wright's formula or Kudo's method (1962) provides estimate of F_A without having any supposition on the genotype structure of the population. On the other hand F_R and hence F_T can only be estimated by the "bioassay" methods. Li and Horvitz (1953) have shown that from a Model II population structure a variety of consistent estimates of F (F_T, in Morton and Yasuda's formula) can be generated. Among the methods of estimation which they describe are methods based upon the (1) total proportion of heterozygotes, (2) product moment correlation between uniting gametes, (3) determinant of the gametic correlation matrix, (4) value of chi-square assuming panmixia in the population, (5) sum of proportions of alleles in homozygous condition among their respective total frequencies, and (6) method of maximum likelihood. When there are only two alleles, the six methods yield identical expressions for F. This is not so when the number of alleles exceeds two. Unfortunately, the sampling variances to be associated with the first five methods of estimation are not known, and thus it is not clear that which, if any, is the method of preference. In the general case, Li and Horvitz were unable to obtain explicit solutions for the gene frequencies and F by the method of maximum likelihood. However for certain special cases (e.g., in case of ABO blood group system) methods are available to obtain efficient estimates of the gene frequencies as well as F, the coefficient of non-randomness (Yasuda, 1968; Schull, Ito and Soni, 1963). In this section we demonstrate the use of parent-offspring correlation, as derived earlier, to provide an alternative estimate* of F. We first give the estimation procedure and once this is done we discuss the advantage of this method over the other existing estimates.

When a character is controlled by two co-dominant alleles A and a at an autosomal locus we have seen through the expression in TABLE 1.4.3 that the parental total measurement (X) and offspring total measurement (Y) have the correlation,

$$\rho_s = \sqrt{\frac{s(1 + 3F)}{(s + 1) + (3s - 1)F}} \qquad (1.8.1)$$

From this one automatically attains

$$\frac{1 + 3F}{1 - F} = \frac{1}{s} \cdot \frac{\rho_s^2}{1 - \rho_s^2} \tag{1.8.2}$$

Now consider a random sample of N families out of which N_s families are with s children each ($s = 1, 2, \ldots, r$) such that

$$\sum_{s=1}^{r} N_s = N$$

Let r_s denote the sample correlation coefficient between parental total score and offspring total score. Then it is easy to see that

$$E\left(\frac{r_s^2}{1 - r_s^2}\right) = \frac{\rho_s^2}{1 - \rho_s^2} + O\left(\frac{1}{N_s}\right) \tag{1.8.3}$$

and $$\text{Var.}\left(\frac{r_s^2}{1 - r_s^2}\right) = \frac{4\,\rho_s^2}{N_s(1 - \rho_s^2)^2} + O\left(\frac{1}{N_s^2}\right) \tag{1.8.4}$$

[for an analogous treatment one can refer to Hetelling (1953) pp. 214].

It is evident now that for large N_s, $\frac{r_s^2}{1 - r_s^2}$ can be taken as a consistent estimate of $\frac{\rho_s^2}{1 - \rho_s^2}$ and consequently one gets consistent estimate of $(1 + 3F)/(1 - F)$ as

$$t_s = \frac{1}{s} \cdot \frac{r_s^2}{1 - r_s^2} \quad \text{for a fixed} \quad s \tag{1.8.5}$$

An estimate of variance of t_s is obviously given by

$$\sigma_s^2 = \frac{4r_s^2}{N_s(1 - r_s^2)^2 s^2} \tag{1.8.6}$$

A pooled estimate of $(1 + 3F)/(1 - F)$ obtained from the whole sample is given by

$$T = \sum_{s=1}^{r} \frac{t_s}{\sigma_s^2} \Big/ \sum_{s=1}^{r} \frac{1}{\sigma_s^2} \tag{1.8.7}$$

and variance of $T = 1/\sum_{1}^{r} \frac{1}{\sigma_s^2}$ (1.8.8)

Note that T is also an consistent estimate of $(1 + 3F)/(1 - F)$ and hence the estimate of F can be written as

$$\hat{F} = \frac{T - 1}{T + 3} \tag{1.8.9}$$

an approximate variance of which is obtained as

$$V(\hat{F}) = \frac{4}{(T + 3)^2 \sum_{s=1}^{r} \frac{1}{\sigma_s^2}} \tag{1.8.10}$$

by using the expressions (1.8.8) and (1.8.9).

A numerical illustration :

Considering the numerical example of section 1.4 (from the family data on MN blood groups analysed by Taylor and Prior (1938) and Race et al. (1942)) we construct the TABLE 1.8.1.

Using the equation (1.8.5) we now have

$$t_1 = 1.0036$$

$$t_2 = 1.1367$$

$$t_3 = 1.9917$$

and $$t_4 = 1.2163$$

TABLE 1.8.1

Parent-offspring correlations for different family size

s	N_s	r_s
1	68	0.7077
2	75	0.8334
3	44	0.9255
4	18	0.9108

Now from (1.8.6) we obtain

$$\frac{1}{\sigma_1^2} = 8.4542$$

$$\frac{1}{\sigma_2^2} = 10.0783$$

$$\frac{1}{\sigma_3^2} = 2.3765$$

and $$\frac{1}{\sigma_4^2} = 2.5232,$$

and thus

$$T = \frac{27.7417}{23.4322} = 1.1839$$

$$\text{Var.}(T) = \frac{1}{23.4322} = 0.0427.$$

Hence, $\hat{F} = \frac{0.1839}{4.1839} = 0.0440$ by (1.8.9) and the standard error of $F = 0.0236$ by (1.8.10).

From the combined sample one also gets 69M, 112MN and 54N individuals among the 235 fathers (the family with serial number 200 is excluded due to the reasons mentioned by the authors). From this, an estimate of F (using any one of the five methods suggested by Li and Horvitz (1953)) is obtained as $F^* = 0.0429$, which is fairly close to the estimate obtained from parent-offspring correlations.

Once the estimation procedure is thus indicated it is natural to ask why one should prefer this method inspite of the existence of simpler ways of estimating it from a random sample of individuals. One of the serious disadvantages of the methods of Li and Horvitz (1953) is that, as described earlier, one cannot have any idea about the standard error of the estimate of F and thus it cannot be decided which one of the five methods is to be adopted to have the most precise estimate. Furthermore though in the two allelic case with codominant genes all of these five methods give identical result but in multi-allelic case situation alters altogether. In that case the estimates obtained by these five methods also differ between themselves.

But in case of the method described here the theoretical expression does not change since the expression of parent-offspring correlation remains the same even in the case of multiple alleles so long as there is no dominance relationship between the alleles concerned. Moreover, the use of parent-offspring correlation enables us to know the standard error of F also.

In case of codominance, Wright had a method of estimating F since in such a case the correlation between mates is given by $2F/(1 + F)$. Use of parent-offspring correlation pays dividend in the sense that the standard error of the estimate of F, thus obtained is much less than that obtained by Wright's method since the former one not only uses the information on mates but also the information from their offspring. Because of these theoretical accounts the use of parent offspring correlation is advocated here to estimate F though from practical view-point of human genetics it is becoming more and more difficult to collect family data than to collect data on unrelated individuals.

CHAPTER II

RESTRICTED RANDOM MATING : A MATING MODEL

2.0 INTRODUCTION

Estimation of gene frequencies for any genetic character involves always directly or indirectly some assumptions about the structure of the population concerned. Furthermore, as we have seen in Chapter I, construction of a model for studying the structure of a population again assumes the prevailing mating scheme in the population. Thus we have seen that repeated random mating leads to a population prescribed by Model I and when a random mating population is inbred to an extent of F $(0 \leq F \leq 1)$, the resulting population behaves like a Model II one. Though these two are the most frequently studied population structures, many geneticists reserve their comments about the applicability of such models. Herein, our contention in this Chapter will be to proceed without any assumption regarding the mating structure at the phenotypic level and later we shall use such a set up to compute gene frequencies from a two generation data. To fix our ideas, we shall first develop the model exclusively for estimating the ABO blood group gene frequencies and later we shall indicate the applicability of such a model (we shall hereafter call it as Restricted random mating model) for the analysis of family data on genetic characters governed by two genes at an autosomal locus having dominant relationship between them.

2.1 DEFINITION OF RESTRICTED RANDOM MATING

In the estimation of gene frequencies, along with the assumptions about the mating systems prevailing in the population, it is also essential to say that the same mating system is operating over generations and the population is in equilibrium so far as the gene frequencies are concerned. In Chapter I we have studied some theorems characterizing the equilibrium conditions for some particular mating schemes. But it is to be noted that in the absence of selection, mutation and migration pressures gene frequencies of a population remain constant from generation to generation whatever be the mating structure prevailing in the population. This fact, apparently, was first noted by Professor J. B. S. Haldane who termed it as "Gene Pool Theorem". However, this idea enables us to search for a estimation procedure which does not need the assumptions regarding the mating structure of the population. In order to do so we assume the phenotypic mating type frequencies to be as it is observed in a sample drawn at random from a population. Thus, in ease of ABO blood groups, there are ten different phenotypic mating types and we assume their relative frequencies to be λ_i's. Note that the conditions on λ_i's are only the natural ones, namely, $0 \leq \lambda_i \leq 1$ and $\Sigma \lambda_i = 1$.

Then we introduce two more parameters, namely, θ and ϕ as

θ = Prob. [an individuals is of genotype AO/his blood group is A]

and ϕ = Prob. [an individual is of genotype BO/his blood group is B].

With these parameters at hand and the general set-up at the phenotypic level we make use of Hardy Weinberg Law only at the level of dividing the general phenotypic mating frequencies into the corresponding genotypic mating frequencies. As for example, consider the mating type A x A. This phenotypic mating can be split up into the three corresponding genotypic matings in the proportions as shown in TABLE 2.1.1.

TABLE 2.1.1

Genotypic mating types given the phenotypic mating A x A and their probabilities

Genotypic mating types	Probability
AO x AO	θ^2
AO x AA	$2\theta(1 - \theta)$
AA x AA	$(1 - \theta)^2$

Thus the relative frequencies of all the 21 genotypic mating types can be written in terms of λ_i's, θ and ϕ. Note that once the estimate of these parameters are obtained, the frequencies of A, B and O genes can be obtained by usual gene count method. The mating system for which the genotypic mating type frequencies are given as shown above will be called in the sequel as Restricted Random Mating, abbreviated as R.R.M.

2.2 USE OF R.R.M. FOR THE ANALYSIS OF ABO BLOOD GROUP DATA

2.2.1 The Model :

From the well accepted theory of mechanism of inheritance of ABO blood groups (whose discussion is avoided here, since it will be discussed in details in Chapter IV) one can write down the parental mating types, their relative frequencies and the conditional distribution of the offspring's blood group as given in TABLE 2.2.1. Note that all the cell probabilities are expressed in terms of the parameters λ_i's, θ and ϕ.

TABLE 2.2.1

Phenotypic mating types, their frequencies and the conditional distribution of the offspring's ABO blood types

Parental Matings		Conditional distribution of the offspring types			
Types	Frequency	O	A	B	AB
O x O	λ_1	1	-	-	-
O x A	λ_2	$\theta/2$	$1 - \frac{\theta}{2}$	-	-
O x B	λ_3	$\phi/2$	-	$1 - \frac{\phi}{2}$	-
O x AB	λ_4	-	$\frac{1}{2}$	$\frac{1}{2}$	--
A x A	λ_5	$\theta^2/4$	$1 - \frac{\theta^2}{4}$	-	-
A x B	λ_6	$\theta\phi/4$	$\frac{\theta\phi}{4} + \frac{\phi(1-\theta)}{4}$	$\frac{\theta\phi}{4} + \frac{\theta(1-\phi)}{4}$	$\frac{\theta\phi}{4} + \frac{1}{2}\theta(1-\phi)$ $+ \frac{1}{2}\phi(1-\theta)$ $+ (1-\theta)(1-\phi)$
A x AB	λ_7	-	$\frac{1}{2}$	$\theta/4$	$\frac{1}{2}(1 - \theta/2)$
B x B	λ_8	$\phi^2/4$	-	$1 - \phi^2/4$	-
B x AB	λ_9	-	$\phi/4$	$\frac{1}{2}$	$\frac{1}{2} - \frac{1}{4}\phi$
AB x AB	λ_{10}	--	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

2.2.2 Estimation of parameters :

To estimate the parameters λ_i's, θ and ϕ, a random sample of families is chosen from the population and the observed frequencies of the parental phenotypic mating types and the phenotypes of the offspring are tabulated in TABLE 2.2.2.

TABLE 2.2.2

Observed frequencies for parental matings and their offspring phenotypes

Parental Matings		Offspring				
Types	Frequencies	0	A	B	AB	Totals
0 x 0	N_1	n_{11}	-	-	-	n_1
0 x A	N_2	n_{21}	n_{22}	-	-	n_2
0 x B	N_3	n_{31}	-	n_{33}	-	n_3
0 x AB	N_4	-	n_{42}	n_{43}	-	n_4
A x A	N_5	n_{51}	n_{52}	-	-	n_5
A x B	N_6	n_{61}	n_{62}	n_{63}	n_{64}	n_6
A x AB	N_7	-	n_{72}	n_{73}	n_{74}	n_7
B x B	N_8	n_{81}	-	n_{83}	-	n_8
B x AB	N_9	-	n_{92}	n_{93}	n_{94}	n_9
AB x AB	N_{10}	-	$n_{10,2}$	$n_{10,3}$	$n_{10,4}$	n_{10}
Totals	N					n

Suppose that the family size distribution is given by Prob. $\{$ a family has k children $\}$ = q_k ; k = 0, 1, ... We shall assume that the family size distribution is the same for all mating types and is also independent of the genotypic frequencies. This is equivalent to saying that with respect to ABO blood groups there is no familial selection operating on the population. This ensures that q_k's are independent of the parameters λ_i's , θ and ϕ .

Now let us have the following notations :

Let T_i denote the mating type with frequency λ_i; i = 1, ..., 10 (e.g., T_5 denotes A x A mating type).

N_{ij} = Number of T_i type of families with j children each. * j = 1, 2, ..., r.

$$N = \sum_{i=1}^{10} N_i \text{ : Total number of families sampled}$$

$$N_i = \sum_{j=1}^{r} N_{ij} \text{ : Total number of } T_i \text{ type of families}$$

$$M_j = \sum_{i=1}^{10} N_{ij} \text{ : Total number of families with } j \text{ children in each}$$

F_{ij}^{l} = l^{th} family of type T_i with j children in it. l = 1, 2, ..., N_{ij}.

n_{ijlk} : number of children in F_{ij}^{l} with blood group k (k = 1, for

0, k = 2 for A, k = 3 for B and k = 4 for AB).

Clearly, $\sum_{k=1}^{4} n_{ijlk} = j$.

p_{ik} : Probability that a child from T_i type of family has blood group k (i = 1, 2, ..., 10; k = 1, 2, 3, 4).

Note that these p_{ik}'s are given in Table 2.2.1 in terms of the parameters θ and ϕ .

Now, Prob.[F_{ij}^{1} has n_{ijlk} children of blood group k; k = 1, 2, 3, 4]

$$= \frac{j!}{\prod_k n_{ijlk}!} \prod_k p_{ik}^{n_{ijlk}}$$

and Prob. [F_{ij}^{1} , ..., $F_{ij}^{N_{ij}}$ have children as observed] *

$$= \prod_{l=1}^{N_{ij}} \left[\frac{j!}{\prod_k n_{ijlk}!} \prod_k p_{ik}^{n_{ijlk}} \right]$$

$$= \frac{(j!)^{N_{ij}}}{\prod_{k,l} n_{ijlk}!} \prod_k p_{ik}^{n_{ij.k}}$$

where, $\sum_{l} n_{ijlk} = n_{ij.k}$ is the total number of children with blood group k in families of T_i with j children in each.

Now, Prob.[N_{ij} families are of size j in N_i families of type T_i; $j = 0, 1, \ldots, r$]

$$= \frac{N_i!}{\prod_{j=0}^{r} N_{ij}!} \prod_{j=0}^{r} q_j^{N_{ij}}$$

Hence, Prob. [N_{ij} families are of size j ($j = 0, 1, \ldots, r$) in N_i families of type T_i and N_{ij} families have children as observed]

$$= \frac{N_i!}{\prod_{j=0}^{r} N_{ij}!} \prod_{j=0}^{r} q_j^{n_{ij}} \frac{\prod_{j=0}^{r} (j!)^{N_{ij}}}{\prod_{j,l,k} n_{ijlk}!} \prod_{k} p_{ik}^{n_{ik}}$$

where, $\sum_{j=0}^{r} n_{ij\cdot k} = n_{ik}$ is the total number of children with blood group k from all families of type T_i.

Furthermore, Prob. [N_i families out of N are of type T_i; $i = 1, 2, \ldots, 10$]

$$= \frac{N!}{\prod_{i=1}^{10} N_i!} \prod_{i=1}^{10} \lambda_i^{N_i}$$

Finally the likelihood of the observed sample is given by

$$L = \frac{N!}{\prod_i N_i!} \prod_i \lambda_i^{N_i} \frac{\prod_i N_i!}{\prod_{i,j} N_{ij}!} \prod_{i,j} q_{i.}^{N_{ij}} \frac{\prod_{i,j} (j\,!)^{N_{ij}}}{\prod_{i,j,l,k} n_{ijlk}\,!} \prod_{i,k} p_{ik}^{n_{ik}}$$

$$= \frac{N!}{\prod_{i,j} N_{ij}!} \cdot \frac{\prod_j (j\,!)^{M_j}}{\prod_{i,j,l,k} n_{ijlk}\,!} \prod_i \lambda_i^{N_i} \prod_j q_j^{M_j} \prod_{i,k} p_{ik}^{n_{ik}} \qquad (2.2.1)$$

where, $M_j = \sum_i N_{ij}$ is the total number of families with j children.

Taking logarithms of both sides we have

$$\log L = \text{Const.} + \sum_i N_i \log \lambda_i + \sum_j M_j \log q_j + \sum_{i,k} n_{ik} \log p_{ik}.$$

From the conditions $\sum_i \lambda_i = 1$, $\sum_j q_j = 1$, $\sum_i N_i = \sum_j M_j = N$,

we easily get the maximum likelihood estimates of λ_i's and q_j's as

$$\hat{\lambda}_i = \frac{N_i}{N} \quad \text{and} \quad \hat{q}_j = \frac{M_j}{N} \qquad (2.2.2)$$

Now all the p_{ik}'s are functions of θ and ϕ alone as given in Table 2.2.1. Suppose we make the transformations $\alpha = \theta/2$ and $\beta = \phi/2$, Table 2.2.1 can be replaced by Table 2.2.3 as

TABLE 2.2.3

Phenotypic mating types, their frequencies and the conditional distribution of the children's ABO blood types

Parental Mating		Conditional Distribution of offspring blood types			
Types	Frequencies	O	A	B	AB
0 x 0	λ_1	1	-	-	-
0 x A	λ_2	α	$1 - \alpha$	-	-
0 x B	λ_3	β	-	$1 - \beta$	-
0 x AB	λ_4	-	$\frac{1}{2}$	$\frac{1}{2}$	-
A x A	λ_5	α^2	$1 - \alpha^2$	-	-
A x B	λ_6	$\alpha\beta$	$\beta(1 - \alpha)$	$\alpha(1 - \beta)$	$(1 - \alpha)(1 - \beta)$
A x AB	λ_7	-	$\frac{1}{2}$	$\alpha/2$	$(1 - \alpha)/2$
B x B	λ_8	β^2	-	$1 - \beta^2$	-
B x AB	λ_9	-	$\beta/2$	$\frac{1}{2}$	$(1 - \beta)/2$
AB x AB	λ_{10}	-	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

From this table we can substitute the values of p_{ik}'s and thus for estimation α and β (and so θ and ϕ) we have the likelihood equations as

$$\frac{\partial \log L}{\partial \alpha} = \frac{C_1}{\alpha} - \frac{C_2}{1 - \alpha} + \frac{C_3}{1 + \alpha} = 0$$

and
$$\frac{\partial \log L}{\partial \beta} = \frac{D_1}{\beta} - \frac{D_2}{1 - \beta} + \frac{D_3}{1 + \beta} = 0.$$

Equivalently,

$$\left.\begin{aligned} &(C_1 + C_2 + C_3)\ \alpha^2 + (C_2 - C_3)\ \alpha - C_1 = 0 \\ \text{and}\quad &(D_1 + D_2 + D_3)\ \beta^2 + (D_2 - D_3)\ \beta - D_1 = 0 \end{aligned}\right\} \qquad (2.2.3)$$

where,

$$C_1 = n_{21} + 2\ n_{51} + n_{61} + n_{63} + n_{73}$$

$$C_2 - C_3 = n_{22} + n_{62} + n_{64} + n_{74}$$

$$C_1 + C_2 + C_3 = n_{21} + 2\ n_{51} + n_{61} + n_{22} + 2\ n_{52} + n_{62} + n_{63} + n_{73} + n_{64} + n_{74}$$

$$D_1 = n_{31} + n_{61} + n_{62} + 2\ n_{81} + n_{92}$$

$$D_2 - D_3 = n_{33} + n_{63} + n_{64} + n_{94}$$

and

$$D_1 + D_2 + D_3 = n_{31} + n_{61} + n_{62} + 2\ n_{81} + n_{92} + n_{33} + n_{63} + n_{64} + 2\ n_{83} + n_{94}\ .$$

Equations in (2.2.3) represent simple quadratic equations which can be solved for α and β . It is also clear that each of the equations has one positive root and one negative root so that the maximum likelihood estimates are uniquely determined as the positive roots of these equations. Thus one has the estimates $\hat{\theta}$ and $\hat{\phi}$ of θ and ϕ as well.

The information matrix corresponding to the sample is computed as follows : First we observe that since λ_i's, q_j's and θ, ϕ are independent parameters, the information matrix will be of the form

$$I = \begin{bmatrix} A & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & C \end{bmatrix} \qquad (2.2.4)$$

where A is the information matrix of order 9 x 9 corresponding to λ_i's, B is the information matrix corresponding to q_j's (of order (r - 1) x (r - 1) and C is the information matrix corresponding to θ and ϕ .

But we may note that

$$A = \begin{bmatrix} \frac{1}{\lambda_1} + \frac{1}{\lambda_{10}} & \frac{1}{\lambda_{10}} & \cdots\cdots & \frac{1}{\lambda_{10}} \\ \frac{1}{\lambda_{10}} & \frac{1}{\lambda_2} + \frac{1}{\lambda_{10}} & \cdots\cdot & \frac{1}{\lambda_{10}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\lambda_{10}} & \frac{1}{\lambda_{10}} & \cdots, & \frac{1}{\lambda_9} + \frac{1}{\lambda_{10}} \end{bmatrix}$$

$$\text{and} \quad B = \begin{bmatrix} \frac{1}{q_1} + \frac{1}{q_r} & \frac{1}{q_r} & \cdots\cdot & \frac{1}{q_r} \\ \frac{1}{q_r} & \frac{1}{q_2} + \frac{1}{q_r} & \cdots\cdot & \frac{1}{q_r} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{q_r} & \frac{1}{q_r} & \cdots\cdot & \frac{1}{q_{r-1}} + \frac{1}{q_r} \end{bmatrix}$$

From these forms of A and B we get their inverses as

$$A^{-1} = \begin{bmatrix} \lambda_1(1-\lambda_1) & -\lambda_1\lambda_2 & \cdots\cdots & -\lambda_1\lambda_9 \\ -\lambda_1\lambda_2 & \lambda_2(1-\lambda_2) & \cdots\cdots & -\lambda_2\lambda_9 \\ \cdots & \cdots & \cdots & \cdots \\ -\lambda_1\lambda_9 & -\lambda_2\lambda_9 & \cdots\cdots & \lambda_9(1-\lambda_9) \end{bmatrix}$$

and similarly B^{-1}.

C is again a diagonal matrix of the form

$$C = \begin{bmatrix} I_{\theta\theta} & 0 \\ 0 & I_{\phi\phi} \end{bmatrix} \qquad (2.2.5)$$

where, $I_{\theta\theta} = \dfrac{\theta(2N_2 + 8N_5 + 2N_6 + N_7) + (4N_2 + 4N_6 + 2N_7)}{2\theta(4-\theta^2)}$

$$I_{\phi\phi} = \frac{(2N_3 + 2N_6 + 8N_8 + N_9) + (4N_3 + 4N_6 + 2N_9)}{2\phi(4-\phi^2)}$$

showing that $\hat{\theta}$ and $\hat{\phi}$ are uncorrelated. This result is intuitively clear also since θ represents the conditional probability that an individuals genotype is AO given that he is of blood group A and ϕ represents the conditional probability that an individual is of genotype BO given that his blood group is B.

Now, inverse of I (given by equation (2.2.4) gives the covariance matrix of the asymptotic distribution of $\hat{\lambda}_i$'s, $\hat{q}_i$'s, $\hat{\theta}$ and $\hat{\varphi}$ which is normal.

Note 2.2.1 : It may be noted that families those are usually surveyed may be of the following types :

(a) families in which both the parents are alive and they have one or more offspring;

(b) families with both the parents alive but without any offspring.

(c) families where only one parent is alive;

(d) families where both the parents are dead;

(e) families in which more than two generations are present.

The estimation procedure discussed earlier includes only families of types (a) and (b). The likelihood function of a sample which includes other types of families can be written down analogously and the analysis can be carried through. If families of type (c) are encountered in the sample, the part of the likelihood function corresponding to such families can be written from the probability distribution which in turn may be obtained from Table 2.2.3 by distinguishing the reciprocal crosses and collecting cells according to the phenotype of one parent. The part of the likelihood function corresponding to families of type (d) can be written considering the marginal distribution of the blood type of the offspring population as obtained from Table 2.2.3. For families of type (e) the generation of the head of the household and his children may be classified into one type (a) or (c). The details of the likelihood functions in these cases and the resulting m. l. estimates are not presented herein since they do not pose any theoretical difficulty.

Note 2.2.2 : The parameters λ_i's, θ and ϕ can also be estimated from a random sample of individuals from the population. The phenotypic relative frequencies are just the weighted column totals of Table 2.2.3. Thus, the sample can be treated as coming from a multinomial distribution and the parameters are estimated by usual procedures. However, these estimates cannot be used for any testing purposes for obvious reasons.

Note 2.2.3 : Though the assumption of any mating system is not made explicitly, the following restriction may be noted : Though a general type of distribution is assumed for the mating types at phenotypic level, one observes that when an individual chooses a mate with blood group A, for example, in the above model, it is assumed that the choice is purely at random between AA and AO. Thus one may object that the hypothesis of Hardy Weinberg Law is not dispensed with all together! However, since there is no easy way of detecting an A individual as AA or AO any model which takes this point into account is going to be only of mathematical interest. So, the assumption of random mating in a restricted sense (at genotypic level only) as outlined above does not reduce the generality of the model.

Estimation of the Phenotypic frequencies :

The phenotypic frequencies of the population are estimated by substituting the estimates $\hat{\lambda}_i$'s, $\hat{e}$ and $\hat{\phi}$ (or $\hat{\lambda}_i$'s, $\hat{\alpha}$ and $\hat{\beta}$) in the expressions P_O, P_A, P_B and P_{AB} (which are just the weighted column totals of Table 2.2.3) given as

$$\left.\begin{aligned} P_O &= \lambda_1 + \alpha\lambda_2 + \beta\lambda_3 + \alpha^2\lambda_5 + \alpha\beta\lambda_6 + \beta^2\lambda_8 \\ P_A &= (1-\alpha)\lambda_2 + \frac{\lambda_4}{2} + (1-\alpha^2)\lambda_5 + \beta(1-\alpha)\lambda_6 + \frac{\lambda_7}{2} + \frac{\beta}{2}\lambda_9 + \frac{\lambda_{10}}{4} \\ P_B &= (1-\beta)\lambda_3 + \frac{\lambda_4}{2} + \alpha(1-\beta)\lambda_6 + \frac{\alpha}{2}\lambda_7 + (1-\beta^2)\lambda_8 \\ &\qquad + \frac{\lambda_9}{2} + \frac{\lambda_{10}}{4} \\ \text{and}\quad P_{AB} &= (1-\alpha)(1-\beta)\lambda_6 + \frac{1-\alpha}{2}\lambda_7 + \frac{1-\beta}{2}\lambda_9 + \frac{\lambda_{10}}{2} \end{aligned}\right\} \quad (2.2.6)$$

Formulae for the variances and covariances of these estimates can be obtained with some algebra. However, since these are no where needed explicitly for our purpose, the details are omitted.

Estimation of the gene frequencies :

Knowing that θ proportions of A individuals and ϕ proportion of B individuals are of genotypes AO and BO respectively, by the usual gene count method we arrive at the gene frequency estimates given by

$$\hat{p} = \frac{2\,\hat{P}_A - \hat{\theta}\,\hat{P}_A + \hat{P}_{AB}}{2}$$

$$\hat{q} = \frac{2\,\hat{F}_B - \hat{\phi}\,\hat{P}_B + \hat{P}_{AB}}{2} \qquad (2.2.7)$$

and $$\hat{r} = \frac{2\,\hat{P}_O + \hat{\theta}\,\hat{P}_A + \hat{\phi}\,\hat{P}_B}{2}$$

where, $\hat{P}_O$, $\hat{P}_A$, $\hat{P}_B$ and $\hat{P}_{AB}$ are the estimated phenotypic relative frequencies. However, the variances and covariances of these estimates are rather complicated. One may derive some approximate expressions for them using Boyd's (1956) method.

2.2.3 Construction of the goodness of fit statistic :

In order to test for goodness of fit or any other hypotheses we first require the asymptotic distribution of the characters studied on the families. It may be recalled that the characteristics observed are the family mating type, family size and the number of children belonging to each blood group O, A, B and AB (designated as 1, 2, 3 and 4) respectively). Thus the random variable observed is a vector variable $\underset{\sim}{X}' = (X_1, \ldots, X_{10}, Y_0, \ldots, Y_r, Z_{1,1,1}, \ldots, Z_{1,1,4}, \ldots, Z_{10,r,1}, \ldots, Z_{10,r,4})$, where X_i's denote the family type, Y_j's denote the family size and $Z_{i,j,k}$ denote the number of children, from a family of type T_i with j children, who are of k^{th} blood type. As for example a family of mating type A x A with 4 children one each of each blood group will be recorded as $X_5 = 1$, $Y_4 = 1$, $Z_{5;4,1} = Z_{5,4,2} = Z_{5,4,3} = Z_{5,4,4} = 1$ and the rests are all zero's. The distribution of $\underset{\sim}{X}$ can easily be obtained from Table 2.2.3 for $N = 1$. The characteristic function of this distribution is given as

$$\mathcal{P}(s_j, t_k, u_{jkl}) = \sum_j \lambda_j e^{is_j} \sum_k q_k e^{it_k} (\sum_l p_{jl} e^{iu_{jkl}})^k \qquad (2.2.8)$$

which follows from Feller (1969).

Hence the characteristic function of the joint distribution of N_j, M_k and n_{jkl}'s is

$$\left[\sum_j \lambda_j e^{is_j} \sum_k q_k e^{it_k} (\sum_l p_{jl} e^{iu_{jkl}})^k \right]^N .$$

Now consider the transformation

$$\left.\begin{array}{l} N'_j = \dfrac{N_i - N\ \lambda_i}{\sqrt{N\ \lambda_j}}, \quad M'_k = \dfrac{M_k - N\ q_k}{\sqrt{N\ q_k}} \\[2ex] \qquad n'_{jkl} = \dfrac{n_{jkl} - k\ N\ \lambda_j\ q_k\ p_{jl}}{\sqrt{k\ N\ \lambda_j\ q_k\ p_{jl}}} \end{array}\right\} \qquad (2.2.9)$$

The characteristic function of these transformed variables is given by

$$\varphi_{N'_j,\ M'_k,\ n'_{jkl}}(\underset{\sim}{t}) = \exp.\Big\{ - i\sqrt{N} \sum_j \sqrt{\lambda_j}\ s_j - i\sqrt{N} \sum_k \sqrt{q_k}\ t_k$$

$$- i\sqrt{N} \sum_{j,k,l} \sqrt{k\ \lambda_j\ q_k\ p_{jl}}\ \ u_{jkl} \Big\}$$

$$\text{x} \Big[\sum_{j,k} \lambda_j\ e^{\frac{i\ s_j}{\sqrt{N\ \lambda_j}}}\ q_k\ e^{\frac{i\ t_k}{\sqrt{N\ q_k}}} \Big(\sum_l p_{jl}\ e^{\frac{i\ u_{jkl}}{\sqrt{k\ n\ \ {}_j q_k p_{jl}}}}\Big)^k \Big]$$

By taking appropriate Taylor's expansions for the exponential terms and ignoring terms of order higher than N^{-1} and simplifying, we see that

$$\log \varphi(\underset{\sim}{t}) \doteq -\tfrac{1}{2} \Big[\sum_j s_j^2 + \sum_k t_k^2 + \sum_{j,k,l} u_{jkl}^2 \Big]$$

$$+ \tfrac{1}{2} \Big[\Big(\sum_j s_j \sqrt{\lambda_j} \Big)^2 + \Big(\sum_k t_k \sqrt{q_k} \Big)^2 \Big]$$

$$- \sum_{j,k,l} s_j u_{jkl} \sqrt{k\, q_k\, p_{jl}} - \sum_{j,k,l} t_k u_{jkl} \sqrt{k\, \lambda_j\, p_{jl}}$$

$$+ \tfrac{1}{2} \Big(\sum_{j,k,l} u_{jkl} \sqrt{k\, \lambda_j\, q_k\, p_{jl}} \Big)$$

$$\Big[2 \sum_j s_j \sqrt{\lambda_j} + 2 \sum_k t_k \sqrt{q_k} + \sum_{j,k,l} u_{jkl} \sqrt{k\, \lambda_j q_k p_{jl}}$$

$$- \tfrac{1}{2} \sum_{j,k} \Big[(k-1) \Big(\sum_l u_{jkl} \sqrt{p_{jl}} \Big)^2 \Big].$$

This shows that the asymptotic distribution of $\underset{\sim}{X}$ is multivariate normal.

Again consider $Z''_{jkl} = \sqrt{k\, q_k}\; Z'_{jkl}$.

Thus, $\log \varphi_{N'_j\,,\; M'_k\,,\; \sqrt{k\, q_k}\; n'_{jkl}} (\underset{\sim}{t})$

$$\doteq -\tfrac{1}{2} \Big[\sum_j s_j^2 + \sum_k t_k^2 + \sum_{j,k,l} k\, q_k\, u_{jkl}^2 \Big]$$

$$+ \tfrac{1}{2} \Big[\Big(\sum_j s_j \sqrt{\lambda_j} \Big)^2 + \Big(\sum_k t_k \sqrt{q_k} \Big)^2 \Big]$$

$$- \sum_{j,k,l} s_j u_{jkl}\, k\, q_k \sqrt{p_{jl}} - \sum_{j,k,l} k \sqrt{q_k}\, t_k u_{jkl} \sqrt{\lambda_j\, p_{jl}}$$

$$+ \tfrac{1}{2} \left(\sum_{j,k,l} k\, q_k\, u_{jkl} \sqrt{\lambda_j\, p_{jl}} \right) \left[2 \sum_j s_j \sqrt{\lambda_j} + 2 \sum_k t_k \sqrt{q_k} + \sum_{j,k,l} k\, q_k\, u_{jkl} \sqrt{\lambda_j\, p_{jl}} \right]$$

$$- \tfrac{1}{2} \sum_{j,k} k(k-1)\, q_k \left(\sum_l u_{jkl} \sqrt{p_{jl}} \right)^2$$

By putting, $u_{jkl} = u_{jl}$ for all k, the resulting expression would be the characteristic function of the joint asymptotic distribution of N_j', M_k' and n_{jl}' where $n_{jl}' = (n_{jl} - N\, m\, \lambda_j\, p_{jl})/\sqrt{N\, \lambda_j\, p_{jl}}$ in which $m = \sum_k k\, q_k$. Denoting $m_2 = \sum_k k^2\, q_k$, the logarithm of the characteristic function is given by

$$\log \varphi_{N_j',\, M_k',\, n_{jl}'}(t) \doteq - \tfrac{1}{2} \left[\sum_j s_j^2 + \sum_k t_k^2 + m \sum_{j,l} u_{jl}^2 \right]$$

$$+ \tfrac{1}{2} \left[\left(\sum_j s_j \sqrt{\lambda_j} \right)^2 + \left(\sum_k t_k \sqrt{q_k} \right)^2 \right] - m \sum_{j,l} s_j\, u_{jl} \sqrt{p_{jl}}$$

$$- \sum_{j,l} u_{jl} \sqrt{\lambda_j p_{jl}} \sum_k k \sqrt{q_k}\, t_k + \tfrac{1}{2} m \left(\sum_{j,l} u_{jl} \sqrt{\lambda_j\, p_{jl}} \right) \times$$

$$\left[2 \sum_j s_j \sqrt{\lambda_j} + 2 \sum_k t_k \sqrt{q_k} + m \sum_{j,l} u_{jl} \sqrt{\lambda_j\, p_{jl}} \right]$$

$$- \tfrac{1}{2} (m_2 - m) \sum_j \left(\sum_l u_{jl} \sqrt{p_{jl}} \right)^2 .$$

The characteristic function of the asymptotic distribution of

N'_j and n'_{jl} can be obtained by just putting $t_k = 0$ for all k in the above expression.

Thus we get,

$$\log \varphi_{N'_j ,\, n'_{jl}} (\underset{\sim}{t}) \doteq -\tfrac{1}{2} \left[\sum_j s_j^2 + m \sum_{j,l} u_{jl}^2 \right] + \tfrac{1}{2} \left(\sum_j s_j \sqrt{\lambda_j} \right)^2$$

$$- m \sum_{j,l} s_j u_{jl} \sqrt{p_{jl}} + \tfrac{1}{2} m \left(\sum_{j,l} u_{jl} \sqrt{\lambda_j p_{jl}} \right) \left[2 \sum_j s_j \sqrt{\lambda_j} + m \sum_{j,l} u_{jl} \sqrt{\lambda_j p_{jl}} \right]$$

$$- \tfrac{1}{2} (m_2 - m) \sum_j \left(\sum_l u_{jl} \sqrt{p_{jl}} \right)^2$$

$$= - \tfrac{1}{2} \left[\sum_j (1 - \lambda_j) s_j^2 - \sum_{j=j'} s_j s_{j'} \sqrt{\lambda_j \lambda_{j'}} + 2m \sum_{j,l} (1 - \lambda_j) s_j u_{jl} \sqrt{p_{jl}} \right.$$

$$- 2m \sum_{j=j',l} s_j u_{j'l} \sqrt{\lambda_j \lambda_{j'} p_{j'l}} + \sum_{j,l} u^2_{jl} \left\{ m + (m_2 - m + m^2 \lambda_j) p_{jl} \right\}$$

$$+ \sum_{j,l=l'} u_{jl} u_{jl'} \sqrt{p_{jl} p_{jl'}} \, (m_2 - m + m^2 \lambda_j)$$

$$\left. - m^2 \sum u_{jl} u_{j'l'} \sqrt{\lambda_j \lambda_{j'} p_{jl} p_{j'l'}} . \right]$$

The last summation includes all combinations but for $j = j'$, $l = l'$.

Thus the variance covariance matrix of the asymptotic distribution of N_j' and n_{jl}' is given by

$$\Lambda = \begin{bmatrix} S & S_1 & S_2 & \cdots & S_j & \cdots & S_{10} \\ S_1' & S_{11} & U_{12} & \cdots & U_{1j} & \cdots & U_{1,10} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ S_j' & U_{j1} & U_{j2} & \cdots & U_{jj} & \cdots & U_{j,10} \\ S_{10}' & U_{10,1} & U_{10,2} & \cdots & U_{10,j} & \cdots & U_{10,10} \end{bmatrix} \qquad (2.2.10)$$

where each of S's and U's are matrices as shown below.

$$S = \begin{bmatrix} (1-\lambda_1) & -\sqrt{\lambda_1 \lambda_2} & \cdots\cdots & -\sqrt{\lambda_1 \lambda_{10}} \\ -\sqrt{\lambda_1 \lambda_2} & (1-\lambda_2) & \cdots & -\sqrt{\lambda_2 \lambda_{10}} \\ \cdots & \cdots & \cdots & \cdots \\ -\sqrt{\lambda_1 \lambda_{10}} & -\sqrt{\lambda_2 \lambda_{10}} & \cdots & (1-\lambda_{10}) \end{bmatrix}$$

$$S_j = \begin{bmatrix} -2m\sqrt{\lambda_1 \lambda_j p_{j1}} & -2m\sqrt{\lambda_2 \lambda_j p_{j2}} & -2m\sqrt{\lambda_3 \lambda_j p_{j3}} & -2m\sqrt{\lambda_4 \lambda_j p_{j4}} \\ \cdots & \cdots & \cdots & \cdots \\ 2m(1-\lambda_j)\sqrt{p_{j1}} & 2m(1-\lambda_j)\sqrt{p_{j2}} & 2m(1-\lambda_j)\sqrt{p_{j3}} & 2m(1-\lambda_j)\sqrt{p_{j4}} \\ \cdots & \cdots & \cdots & \cdots \\ -2m\sqrt{10\,\lambda_j p_{j1}} & -2m\sqrt{\lambda_{10} \lambda_j p_{j2}} & -2m\sqrt{\lambda_{10} \lambda_j p_{j3}} & -2m\sqrt{\lambda_{10} \lambda_j p_{j4}} \end{bmatrix}$$

for $j = 1, 2, \dots. 10$.

$$U_{jj} = \begin{bmatrix} m + \lambda_j p_{j1} & \alpha_j \sqrt{p_{j1} p_{j2}} & \alpha_j \sqrt{p_{j1} p_{j3}} & \alpha_j \sqrt{p_{j1} p_{j4}} \\ \alpha_j \sqrt{p_{j1} p_{j2}} & m + \alpha_j p_{j2} & \alpha_j \sqrt{p_{j2} p_{j3}} & \alpha_j \sqrt{p_{j2} p_{j4}} \\ \alpha_j \sqrt{p_{j1} p_{j3}} & \alpha_j \sqrt{p_{j2} p_{j3}} & m + \alpha_j p_{j3} & \alpha_j \sqrt{p_{j3} p_{j4}} \\ \alpha_j \sqrt{p_{j1} p_{j4}} & \alpha_j \sqrt{p_{j2} p_{j4}} & \alpha_j \sqrt{p_{j3} p_{j4}} & m + \alpha_j p_{j4} \end{bmatrix}$$

for $j = 1, 2, \dots, 10$ where $\alpha_j = m_2 - m + m^2 \lambda_j$.

$$U_{jk} = (-m^2 \sqrt{\lambda_j \lambda_k}) \begin{bmatrix} \sqrt{p_{j1} p_{k1}} & \sqrt{p_{j1} p_{k2}} & \sqrt{p_{j1} p_{k3}} & \sqrt{p_{j1} p_{k4}} \\ \sqrt{p_{j2} p_{k1}} & \sqrt{p_{j2} p_{k2}} & \sqrt{p_{j2} p_{k3}} & \sqrt{p_{j2} p_{k4}} \\ \sqrt{p_{j3} p_{k1}} & \sqrt{p_{j3} p_{k2}} & \sqrt{p_{j3} p_{k3}} & \sqrt{p_{j3} p_{k4}} \\ \sqrt{p_{j4} p_{k1}} & \sqrt{p_{j4} p_{k2}} & \sqrt{p_{j4} p_{k3}} & \sqrt{p_{j4} p_{k4}} \end{bmatrix}$$

for $j \neq k = 1, 2, \dots, 10$ and $U_{jk} = U'_{kj}$.

The rank of the Λ matrix is $(10 - 1) + (24 - 9) = 24$. In fact by observing that many of the p_{jk}'s are zero's, we can reduce the order of this matrix from 50×50 to 34×34. Then, note that there is one linear constraint in the λ_i's and one linear constraint in each

p_{jk}'s for $j = 2, \ldots, 10$. From this considerations one observes that the rank of Λ is 24.

The reduced Λ matrix of order 34 x 34, being a positive semi-definite one of rank 24, will have 24 positive eigen values and the rest 10 eigen values are equal to zero. If Q denotes the matrix of the corresponding 24 eigen vectors (and thus Q is a matrix of order 24 x 34), we have

$$Q \Lambda Q' = \Delta$$

where Δ is a diagonal matrix of order 24 x 24. In fact Δ is the matrix of the positive eigen values. It is easy to note that the random variable $\underset{\sim}{Y} = Q \underset{\sim}{X}$ follows asymptotically a 24-dimensional multivariate normal distribution with mean vector $\underset{\sim}{0}$ and variance covariance matrix $Q \Lambda Q' = \Delta$ (Rao, 1965).

Hence, $Z = \underset{\sim}{Y}' \nabla \underset{\sim}{Y}$ is asymptotically distributed as a chi-square with 24 degrees of freedom (∇, being the inverse of Δ, is a diagonal matrix whose diagonal elements are the reciprocal of the positive eigen values of Δ).

This Z is the goodness of fit statistic which is looked for.

2.2.4 A numerical example :

As an illustration to the theory developed in the earlier three sections we present the data on ABO blood groups from families in some villages of Purulia district of West Bengal, India. The observations on parental mating types and the distribution of blood groups in offspring are tabulated in Table 2.2.4.

TABLE 2.2.4

Mating types and distribution of blood groups in offspring

Mating		Children				
Type	Frequency	O	A	B	AB	Total
0 x 0	24	40	-	-	-	40
0 x A	126	139	177	-	-	316
0 x B	63	37	-	100	-	137
0 x AB	4	-	2	8	2	12
A x A	87	85	118	-	-	203
A x B	53	2	2	2	65	71
A x AB	9	-	3	3	6	12
B x B	79	22	-	157	-	179
B x AB	10	-	-	3	4	7
AB x AB	4	-	-	-	2	2
Total	459	325	302	273	79	979

From the expressions in (2.2.2) we at once get the estimates of λ_i's as

$$\left.\begin{array}{lll} \hat{\lambda}_1 = 0.0523, & \hat{\lambda}_2 = 0.2745, & \hat{\lambda}_3 = 0.1373 \\ \hat{\lambda}_4 = 0.0087, & \hat{\lambda}_5 = 0.1895, & \hat{\lambda}_6 = 0.1155 \\ \hat{\lambda}_7 = 0.0196, & \hat{\lambda}_8 = 0.1721, & \hat{\lambda}_9 = 0.0218 \\ & \text{and} \quad \hat{\lambda}_{10} = 0.0087 & \end{array}\right\} \qquad (2.2.11)$$

Equations (2.2.3) take the form

$$802 \ \alpha^2 + 250 \ \alpha - 316 = 0$$

$$\text{and} \quad 570 \ \beta^2 + 171 \ \beta - .85 = 0$$

from which the admissible solutions are obtained as

$$\hat{\alpha} = 0.4909 \quad \text{and} \quad \hat{\beta} = 0.2643 \qquad (2.2.12)$$

Hence, the estimates of θ and ϕ are given by

$$\hat{\theta} = 2\,\hat{\alpha} = 0.9818 \quad \text{and} \quad \hat{\phi} = 2\,\hat{\beta} = 0.5286 \qquad (2.2.13)$$

Feeding (2.2.11) and (2.2.12) into the equations of (2.2.6) we obtain

$$\hat{P}_0 = 0.2960, \quad \hat{P}_A = 0.3183, \quad \hat{P}_B = 0.3251 \quad \text{and} \quad \hat{P}_{AB} = 0.0606$$

and thus from the expressions in (2.2.7) we obtain the estimates of the gene frequencies as

$$\hat{p} = 0.192370, \quad \hat{q} = 0.269458 \quad \text{and} \quad \hat{r} = 0.538172.$$

At this stage it is interesting to note that taking $\hat{\theta}$ and $\hat{\phi}$ as given by (2.2.13) we can breakdown the phenotypic frequencies of the parental as well as the offspring generation (which are in turn obtained from Table 2.2.4) to get the corresponding genotypic frequencies of these two populations. These are shown in the Table below :

TABLE 2.2.5

Genotypic blood group frequencies for parental and offspring generations

Genotypes	Parental Population	Offspring Population
OO	241.0000	325.0000
OA	355.4116	296.5036
AA	6.5884	5.4964
OB	150.0940	144.2805
BB	133.9060	128.7195
AB	31.0000	79.0000
Total	918.0000	979.0000

From this table we get the estimates of the gene frequencies for these two populations as

	Parental Population	Offspring Population
p	0.217641	0.197393
q	0.244502	0.245516
r	0.537857	0.557091

From this table one can easily notice that apparently the gene frequencies in these two generations are not significantly different from one another which is expected because of the gene pool theorem of Professor J. B. S. Haldane. Of course the statement remains valid if only we assume the selection, mutation and migration pressures to be inoperative at the ABO locus for this population.

So long, we did not consider the variances and covariances of the estimates. From expression (2.2.5) we note that $\hat{\theta}$ and $\hat{\phi}$ are uncorrelated and

$$V(\hat{\theta}) = 1/I_{\theta\theta} = 0.003354$$

and $$V(\hat{\phi}) = 1/I_{\phi\phi} = 0.000417.$$

The other variances and covariances are also obtained similarly from the corresponding expressions.

We are deliberately omitting the computation of goodness of fit statistic since it is only a matter of obtaining the eigen values and the eigen vectors of the Λ matrix (of order 34 x 34) which can easily be done by using an electronic computer. However, for getting the Λ matrix and hence the goodness of fit statistic we need the two statistics of the family size distribution given by

$$m = \Sigma k \, q_k = 2.1285$$

and $$m_2 = \Sigma k^2 \, q_k = 7.2702$$

The goodness of fit statistic, for an analogous problem, will, however, be presented with illustration in the next section.

2.3 USE OF R.R.M. IN TWO ALLELIC CASES

The main object in this section will be to develop a procedure for testing whether any particular population is in equilibrium incorporating the model of restricted random mating. We shall do this for an autosomal character governed by two alleles, one of which being dominant over the other. For such an autosomal character we shall illustrate the computation of the goodness of fit chi-square statistic whose construction is already indicated in the earlier section.

2.3.1 Two alleles at an autosomal locus : Goodness of fit statistic :

To start with, let us consider two alleles A and a at an autosomal locus. Let us also assume that A is dominant over a so that only two phenotypes $\bar{A}$ (representing genotypes AA and Aa collectively) and $\bar{a}$ (representing the genotype aa) are distinguishable. Let θ be the conditional probability that an individual is of genotype Aa given that his phenotype is known to be $\bar{A}$. If λ_1, λ_2, λ_3 represent the probabilities of the $\bar{A} \times \bar{A}$, $\bar{A} \times \bar{a}$ and $\bar{a} \times \bar{a}$ mating types respectively then we have the conditional distribution of the offspring phenotypes as given by Table 2.3.1.

TABLE 2.3.1

Parental matings and the offspring phenotype probabilities

Parental Matings		Offspring	
Type	Frequency	$\bar{A}$	$\bar{a}$
$\bar{A} \times \bar{A}$	λ_1	$1 - \alpha^2$	α^2
$\bar{A} \times \bar{a}$	λ_2	$1 - \alpha$	α
$\bar{a} \times \bar{a}$	λ_3	0	1

where $\alpha = \theta/2$.

Now, in a random sample of N families chosen from a population let the observed frequencies of the parental mating types and the phenotypes of the offspring be as tabulated in Table 2.3.2.

TABLE 2.3.2

Observed frequencies for parental matings and their offspring phenotypes

Parental Matings		Offspring		
Type	Frequency	$\bar{A}$	$\bar{a}$	Total
$\bar{A}$ x $\bar{A}$	N_1	n_{11}	n_{12}	n_1
$\bar{A}$ x $\bar{a}$	N_2	n_{21}	n_{22}	n_2
$\bar{a}$ x $\bar{a}$	N_3	-	n_{32}	n_3
Total	N			n

The logarithm of the likelihood of such a sample is given by

$$\log L = \text{Const.} + \sum_i N_i \log \lambda_i + \sum_j M_j \log q_j + \sum_{i,k} n_{ik} \log p_{ik} \qquad (2.3.1)$$

[the notations used here are same as those used for the derivation of the expression in (2.2.1)].

We, now, automatically get the maximum likelihood estimates of λ_i's and α as

$$\hat{\lambda}_i = N_i/N \qquad (2.3.2)$$

and $\hat{\alpha}$ is the positive root of the equation

$$C_1 \alpha^2 + C_2 \alpha - C_3 = 0 \tag{2.3.3}$$

where $C_1 = 2 n_{11} + 2 n_{12} + n_{21} + n_{22}$

$C_2 = n_{21}$

and $C_3 = 2 n_{12} + n_{22}$.

Thus, the maximum likelihood estimate of θ is given by $\hat{\theta} = 2 \hat{\alpha}$.

It is easy to see that the asymptotic variances and covariances of the estimates of λ_i's can be estimated by

$$\left.\begin{aligned} V(\hat{\lambda}_1) &= \frac{N_1(N - N_1)}{N^3} \\ V(\hat{\lambda}_2) &= \frac{N_2(N - N_2)}{N^3} \\ \text{Cov.}(\hat{\lambda}_1, \hat{\lambda}_2) &= -\frac{N_1 N_2}{N^3} \\ \text{and} \quad V(\hat{\lambda}_3) &= V(1 - \hat{\lambda}_1 - \hat{\lambda}_2) \\ &= V(\hat{\lambda}_1) + V(\hat{\lambda}_2) + 2\,\text{Cov.}(\hat{\lambda}_1, \hat{\lambda}_2) \\ &= \frac{N_3(N - N_3)}{N^3} \end{aligned}\right\} \tag{2.3.4}$$

In order to get the asymptotic variance of θ we first have the information about α from the whole data as

$$I(\alpha) = \frac{n_2 + \alpha(4 n_1 - n_2)}{\alpha(1 - \alpha^2)} \tag{2.3.5}$$

Hence, the asymptotic variance of $\hat{\alpha}$

$$V(\hat{\alpha}) = \frac{\alpha(1 - \alpha^2)}{n_2 + \alpha(4 n_1 - n_2)} \tag{2.3.6}$$

and thus $V(\hat{\theta}) = 4\, V(\hat{\alpha})$.

Before proceeding to the testing of equilibrium let us now consider an illustration and compute the goodness of fit chi-square statistic. For this purpose, let us consider again the data collected by Race et al (1948, 1949) on the S-factor of MNSs blood groups. When s- antiserum is not used we can distinguish only two phenotypes namely S- (including genotypes SS and Ss) and s- (including the genotype ss). The observed mating type frequencies and the phenotypic frequencies of the children are shown in Table 2.3.3.

TABLE 2.3.3

The S group of 123 families with 293 children

Parental Mating		Children		
Types	Frequencies	S-	s-	Total
S- x S-	44	92	12	104
S- x s-	57	84	52	136
s- x s-	22	-	53	53
Total	123	176	117	293

Estimates of the mating type frequencies are seen to be

$$\hat{\lambda}_1 = 0.3577, \quad \hat{\lambda}_2 = 0.4634 \quad \text{and} \quad \hat{\lambda}_3 = 0.1789$$

Note that $C_1 = 344$, $C_2 = 84$ and $C_3 = 76$ and thus $\hat{\alpha}$ is the positive root of

$$344 \ \alpha^2 + 84 \ \alpha - 76 = 0$$

which gives $\hat{\alpha} = 0.3635$ and thus $\hat{\theta} = 2 \ \hat{\alpha} = 0.7270$.

The variances and covariances of the estimates are obtained as

$$V(\hat{\lambda}_1) = 0.001868, \quad V(\hat{\lambda}_2) = 0.002022$$

$$\text{and} \quad \text{Cov.}(\hat{\lambda}_1, \hat{\lambda}_2) = -0.001348$$

$$V(\hat{\theta}) = 0.001327.$$

Now, a derivation analogous to that of (2.2.10) gives the variance covariance matrix of $\underset{\sim}{X}' = (N_1', N_2', N_3', n_{11}', n_{12}', n_{21}', n_{22}', n_{32}')$ where $N_i' = \dfrac{M_i - N \ \lambda_i}{\sqrt{N. \ \lambda_i}}$; $i = 1, 2, 3$

$$n_{ij}' = \frac{n_{ij} - N.m.\lambda_i \cdot p_{ij}}{\sqrt{N \ \lambda_i \ p_{ij}}}$$

as

$$\Lambda = \begin{bmatrix} S & S_1 & S_2 & S_3 \\ S_1' & U_{11} & U_{12} & U_{13} \\ S_2' & U_{12}' & U_{22} & U_{23} \\ S_3' & U_{13}' & U_{23}' & U_{33} \end{bmatrix} \qquad (2.3.7)$$

where

$$S = \begin{bmatrix} 1 - \lambda_1 & -\sqrt{\lambda_1 \lambda_2} & -\sqrt{\lambda_1 \lambda_3} \\ -\sqrt{\lambda_1 \lambda_2} & 1 - \lambda_2 & -\sqrt{\lambda_2 \lambda_3} \\ -\sqrt{\lambda_1 \lambda_3} & -\sqrt{\lambda_2 \lambda_3} & 1 - \lambda_3 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 2m(1 - \lambda_1)\sqrt{p_{11}} & 2m(1 - \lambda_1)\sqrt{p_{12}} \\ -2m\sqrt{\lambda_1 \lambda_2 p_{11}} & -2m\sqrt{\lambda_1 \lambda_2 p_{12}} \\ -2m\sqrt{\lambda_1 \lambda_3 p_{11}} & -2m\sqrt{\lambda_1 \lambda_3 p_{12}} \end{bmatrix}$$

$$S_2 = \begin{bmatrix} -2m\sqrt{\lambda_1 \lambda_2 p_{21}} & -2m\sqrt{\lambda_1 \lambda_2 p_{22}} \\ 2m(1 - \lambda_2)\sqrt{p_{21}} & 2m(1 - \lambda_2)\sqrt{p_{22}} \\ -2m\sqrt{\lambda_2 \lambda_3 p_{21}} & -2m\sqrt{\lambda_2 \lambda_3 p_{22}} \end{bmatrix}$$

$$S_3 = \begin{bmatrix} -2m\sqrt{\lambda_1 \lambda_3} \\ -2m\sqrt{\lambda_2 \lambda_3} \\ 2m(1 - \lambda_3) \end{bmatrix}$$

$$U_{11} = \begin{bmatrix} m + \lambda_1 p_{11} & \lambda_1 \sqrt{p_{11} p_{12}} \\ \lambda_1 \sqrt{p_{11} p_{12}} & m + \lambda_1 p_{12} \end{bmatrix}$$

$$U_{22} = \begin{bmatrix} m + \lambda_2 p_{21} & \lambda_2 \sqrt{p_{21} p_{22}} \\ \lambda_2 \sqrt{p_{21} p_{22}} & \alpha m + \alpha_2 p_{22} \end{bmatrix}$$

$$U_{33} = m + \alpha_3$$

$$U_{12} = \begin{bmatrix} \sqrt{p_{11} p_{21}} & \sqrt{p_{11} p_{22}} \\ \sqrt{p_{12} p_{21}} & \sqrt{p_{12} p_{22}} \end{bmatrix} (-m^2 \sqrt{\lambda_1 \lambda_2})$$

and $$U_{23} = (-m^2 \sqrt{\lambda_2 \lambda_3}) \begin{bmatrix} \sqrt{p_{11}} \\ \sqrt{p_{22}} \end{bmatrix}$$

Note that two statistics m and m_2 (m_2 appears in α_j's given by $\alpha_j = m_2 - m + m^2 \lambda_j$) regarding the family size distribution appear here. From the family data of Race et al. We have

$$m = \Sigma\, k\, q_k = 2.3821$$

and
$$m_2 = \Sigma\, k^2\, q_k = 7.1463.$$

The Λ matrix, thus computed is shown in the next page. It is easy to see that the matrix is of rank 5 since there is one linear constraint with λ_i's $(\lambda_1 + \lambda_2 + \lambda_3 = 1)$ and two with p_{ij}'s $(p_{11} + p_{12} = 1 = p_{21} + p_{22})$. The 5 positive eigen values are given by 0.8143, 1.1360, 5.5816, 14.0481 and 12.2978. The corresponding matrix of eigen vectors (taken as column vectors), Q', is given by

$$Q' = \begin{bmatrix} 0.6800 & -0.3064 & 0.0029 & -0.2457 & -0.1599 \\ -0.6371 & -0.2421 & -0.0680 & 0.2471 & -0.0802 \\ 0.0637 & 0.8229 & 0.1054 & -0.0503 & 0.3552 \\ -0.2624 & 0.0959 & 0.5566 & -0.5657 & -0.4204 \\ -0.0948 & 0.0347 & 0.2010 & -0.2043 & -0.1518 \\ 0.1727 & 0.0537 & 0.4754 & 0.5582 & -0.2174 \\ 0.1359 & 0.0422 & 0.3740 & 0.4392 & -0.1710 \\ -0.0384 & -0.3940 & 0.5178 & -0.0983 & 0.7520 \end{bmatrix}$$

Variance Covariance Matrix of the Normalised observation vector (Λ matrix)

0.6423	-0.4072	-0.2530	1.8780	1.0395	-1.5246	-1.1996	-1.2053
	0.5366	-0.2879	-1.8246	-0.6590	1.0091	1.5808	-1.3717
		0.8211	-1.1336	-0.4094	-1.0780	-0.8482	1.9121
			8.3923	2.1707	-1.7078	-1.3438	-1.3502
				3.1661	-0.6168	-0.4854	-0.4877
					6.9489	3.5932	-1.2840
						5.2091	-1.0103
							8.1613

Note that the observed transformed vector is of the form

$$\underset{\sim}{X}' = (-0.000008,\ 0.000010,\ -0.000006,\ -0.112633,\ -0.041647,\ 0.022883,\ 0.017977,\ 0.126521).$$

Hence, $\underset{\sim}{Y} = Q\,\underset{\sim}{X}$ is given by

$$\underset{\sim}{Y}' = (0.0350,\ -0.0601,\ 0.0121,\ 0.0805,\ 0.1408).$$

From the theory developed in section 2.2.3 we know $\underset{\sim}{Y}$ follows asymptotically a multivariate normal distribution with variance covariance matrix $Q \wedge Q' = \Delta$ which is, in fact, a diagonal matrix whose diagonal elements are the positive eigen values of $\wedge$. Denoting them by $d_1, \ldots, d_5$ we have, again, from the theory as developed in section 2.2.3

$$Z = Y' \nabla Y = \sum_{i=1}^{5} y_i^2 \,/\, d_i$$

follows a χ^2 distribution with 5 d.f.

In the present example, $Z = 0.0069$ which is far from being significant ($P > 0.99$). Thus, the data seem to be in good agreement with the model considered here.

2.3.2 Equilibrium condition and a proposed test criterian :

From theorem 1.1.1 (in Chapter I) we have seen that a population will be in equilibrium with respect to an autosomal locus with two alleles A and a if and only if the Aa x Aa matings are twice as frequent as those between the two different homozygotes (AA x aa and aa x AA). Thus, in the present set-up, we have the equilibrium condition as

$$\frac{\lambda_1}{\lambda_2} = \frac{2(1 - \theta)}{\theta^2} \qquad (2.3.8)$$

Now with $\alpha = \theta/2$, we have from (2.3.8),

$$\lambda_1 = \frac{(1 - 2\alpha)\lambda_2}{2\alpha^2}$$

$$\text{and} \quad \lambda_3 = 1 - \lambda_1 - \lambda_2 = \frac{2\alpha^2 - \lambda_2[\alpha^2 + (1 - \alpha)^2]}{2\alpha^2}$$

Thus, in this case, there are only two independent parameters, namely α and λ_2. Inserting these values of λ_1 and λ_3 in (2.3.1) we have the maximum likelihood estimates of the parameters as :

α is the positive root of $A\alpha^5 + B\alpha^4 + C\alpha^3 + D\alpha^2 + E\alpha + F = 0$.

where, $A = 8\ n_{11}$

$$B = 8\ n_{12} + 4\ n_{21} + 4\ n_{22} - 12\ n_{11} - 12\ N_2$$

$$C = 28\ N_2 + 8\ n_{11} - 20\ n_{12} - 6\ n_{21} - 10\ n_{22}$$

$$D = N_1 - 25\ N_2 - 2\ n_{11} + 20\ n_{12} + 4\ n_{21} + 10\ n_{22}$$

$$E = -N_1 + 11\ N_2 - 10\ n_{12} - n_{21} - 5\ n_{22}$$

and $F = -2\ N_2 + 2\ n_{12} + n_{22}$

$$\hat{\lambda}_2 = \frac{N_1 + N_2}{N} \cdot \frac{2\ \hat{\alpha}^2}{\hat{\alpha}^2 + (1 - \hat{\alpha})^2}$$

$$\hat{\lambda}_1 = \frac{(1 - 2\hat{\alpha})\ \hat{\lambda}_2}{2\ \hat{\alpha}^2}$$

and
$$\hat{\lambda}_3 = \frac{2\ \hat{\alpha}^2 - \hat{\lambda}_2 \left[\hat{\alpha}^2 + (1 - \hat{\alpha})^2 \right]}{2\ \hat{\alpha}^2}$$

Considering the same example, the estimates obtained are as follows :

$$\hat{\alpha} = 0.3814 \quad \text{and hence} \quad \hat{\theta} = 2\,\hat{\alpha} = 0.7628$$

$$\hat{\lambda}_2 = 0.4416\ , \quad \hat{\lambda}_1 = 3600 \quad \text{and} \quad \hat{\lambda}_3 = 0.1984.$$

The variance covariance matrix of these estimates can easily be obtained by computing the information matrix and then by inverting it. However, the details are left out since it does not present any theoretical difficulty.

At this stage, let us recall that in the general case there are three linearly independent parameters (namely, λ_1, λ_2 and θ) whereas under equilibrium assumption they are only two in number (namely, λ_2 and θ). Now, let us denote the estimates (in the expression (2.3.9)) with subscript e (to denote that they are obtained under the assumption of equilibrium). Thus we have

$$\text{Max. } \log L = \log L_e = \text{Const.} + \sum_i N_i \log \hat{\lambda}_{ie} + \sum_i \sum_j n_{ij} \log \hat{p}_{ije}$$

[under equilibrium]

whereas, in general

$$\text{Max. } \log L = \log L_g = \text{Const.} + \sum_i N_i \log \hat{\lambda}_{ig} + \sum_i \sum_j n_{ij} \log \hat{p}_{ijg}$$

[the estimates with subscript g are obtained from (2.3.2) and (2.3.3)].

Now, we can easily write down the likelihood ratio test criterion for testing the hypothesis of equilibrium as

$$\begin{aligned} -2 \log \Lambda &= 2 \left[\log L_g - \log L_e \right] \\ &= 2 \sum_i N_i \left[\log \hat{\lambda}_{ig} - \log \hat{\lambda}_{ie} \right] \\ &\quad + 2 \sum_i \sum_j n_{ij} \left[\log \hat{p}_{ijg} - \log \hat{p}_{ije} \right] \end{aligned}$$

$$= 2 \sum_i N_i \log \frac{\lambda_{ig}}{\lambda_{ie}} + 2\, n_{11} \log \frac{1 + \hat{\alpha}_g}{1 + \hat{\alpha}_e}$$

$$+ 2(n_{11} + n_{21}) \log \frac{1 - \hat{\alpha}_g}{1 - \hat{\alpha}_e} + 2(2n_{12} + n_{22}) \log \frac{\hat{\alpha}_g}{\hat{\alpha}_e}$$

which follows asymptotically a χ^2 - distribution with 3 - 2 = 1 d.f.

For the present example the computed value of the chi-square statistic is 0.6855 which is, again, far from being significant ($P > 0.40$) which indicates that the population is under equilibrium with respect to the locus of S blood group factor.

CHAPTER - III

SOME STATISTICAL MODELS FOR HUMAN MULTIPLE BIRTHS

3.0 INTRODUCTION

In experimental animals or plants it is relatively easy to study the interactions of nature and nurture by controlling environment at will in which the phenotypic properties develop. In case of human genetics, in general, though it is not feasible, nevertheless certain phenomena in man approach the ideal arrangements of experimental design. The most significant of these being the multiple births. Experiments to analyse the effect of a spectrum of environments can be designed by including the 'identical' twins which are isogenic, whence the 'non-identical' (fraternal) twins provide scope to study the effects of different genotypes under identical environmental conditions. In this chapter we are going to review briefly the biological implications of human multiple births and then we shall switch on to construct some models to lay down the probabilities of these type of births. It will be observed that some of these models extend the already existing ones (Das, 1953-56; Bulmer, 1958; Allen, 1960; etc.) taking into account the limitations involved there.

3.1 BIOLOGY OF TWINS

At least five distinctly different biological situations can be postulated, giving rise to twins, namely : the proliferation and fertilization of (i) two ova; (ii) only one ovum, when this fertilized egg subsequently goes through a scission at some early stage of its development; (iii) a binucleate egg which subsequently divides; (iv) an egg and a large polar body; and (v) the scission products of an egg which has divided prior to fertilization. Though a direct cytological evidence is lacking to demonstrate that all these possibilities are met in case of human but serologic and somatologic evidences indicate that at least the first two of the possibilities do occur and with a good amount of regularity. For our purpose we restrict to these two common situations only.

So, there are two different types of twins; 'identical' or monozygotic (MZ) and 'non-identical' or dizygotic (DZ). As the terminology suggests, the former comes into existence by a twinning division of a single zygote, formed by the fertilization of a single ovum by a single mature sperm and the latter type of twins are formed as a result of the two independently liberated ova by two separate mature sperms. They may, therefore, have like or unlike sexes and are generally no more similar than two sibs born of the same parents. On the other hand the members of a MZ twin pair must, of necessity, be of the same sex and are of identical genotypes (isogenic).

The scission of a single zygote which occurs at an early stage of its life (Dahlberg, 1926) may, in some cases, be satisfied after only one such occurance i.e. after the formation of a pair of twins, but it may persist even further and the one, or the other, or both of the members of the scissioned zygotes may again undergo similar divisions producing MZ triplets, quadruplets, etc. Thus, as Das (1953) writes, "in order that a monozygotic twin pair should result, the scission must take place once and only once. In order that a monozygotic triplet should result, the splitting must take place twice : the first division of the single zygote produces an identical pair, one and only one member of which should then similarly split into two, but once". Similar visualisation can be carried out for the formation of higher monozygotic multiplicity. But the actual phenomenon that takes place is far more complicated since 'death of a zygote' is not a less important event than the birth of the same.

The genesis of DZ twins is altogether different from that of the MZ twins. Two ova liberated from one, or both of the ovaries within a small interval of time and fertilised by two independent sperms lead to the formation of two zygotes which develop and grow simultaneously and thus form a pair of DZ twins. Thus, they may be either like-sexed or unlike-sexed.

Therefore, as Dahlberg (1926) and Greulich (1935) observed, the two types of twins are the manifestations of two basically distinct phenomena and are not merely the different expressions of one and the same twinning tendency. These two basic phenomena, in different sequences can lead to, at least theoretically, multiple births of any order. As for example, a r-zygous n-tuple birth occurs when to start with altogether r ova are released and after fertilization in all exactly (n - r) scissions take place to them. Again prenatal mortality factors are altogether ignored here.

3.2 MATERNAL AGE, PARITY OF BIRTH AND TWINNING RATES

Both the over-all frequency of twinning and the frequency of each of the two types of twins vary considerably from country to country. As an example one can have a range of this rate as 1 in

1000 maternities (among the Annamese of Cochin China; Newman, 1940) to as many as 35 in 1000 maternities in South Rhodesia (Ross, 1952). But merely this racial difference in twinning rates hardly throws any light in understanding the twinning meohanism. But as already observed by a number of authors it is seen that there exists an interesting relationship between the age of mother and rate of twin confinements. Enders and Stern (1948), with their data on American Whites and Negroes, have drawn the following conclusions :

(i) The chance of a mother having DZ twin confinements increases steadily with the age of the mother upto a certain age group and then abruptly comes down.

(ii) The chance for the incidence of MZ twins shows a slight but steady tendency to increase with the age of the mother throughout the whole reproductive range of age.

Sarkar (1944) collected some Indian data and proved a similar relationship between the frequency of unlike-sexed twins and the age of mothers in case of 84 twins. The conclusion (i) of Enders and Stern is confirmed by some later studies (Waterhouse, 1950; Stocks, 1952; Bulmer, 1958; etc.). We present here (Table 3.2.1 and Fig. 3.2.1) the data of Stocks for England and Wales which show that the proportion of DZ twin maternities increases with the age of mother from puberty to a peak in the age group 35-39, falling thereafter rather more steeply than it rose. The incidence of MZ twins, on the other hand, is virtually independent of maternal age, rising from 3.05 to 4.29 per thousand.

TABLE 3.2.1

Frequency of MZ and DZ twins according to the age of mother
(After Stocks, 1952)

Age of mother (in years)		Twinning rate per 1000 maternities	
Age group	Average age of all mothers	MZ	DZ
Under 20	19.0	3.05	3.30
20 - 24	22.8	3.23	5.26
25 - 29	27.5	3.31	7.91
30 - 34	32.3	3.51	10.82
35 - 39	37.2	3.86	12.79
40 - 44	41.8	3.55	9.47
45 and abcve	46.3	4.29	2.61

It is at once apparent that the comparison of gross rates of twin production is invalid unless accompanied by the distribution of maternal age of all births; for a relation of the kind depicted

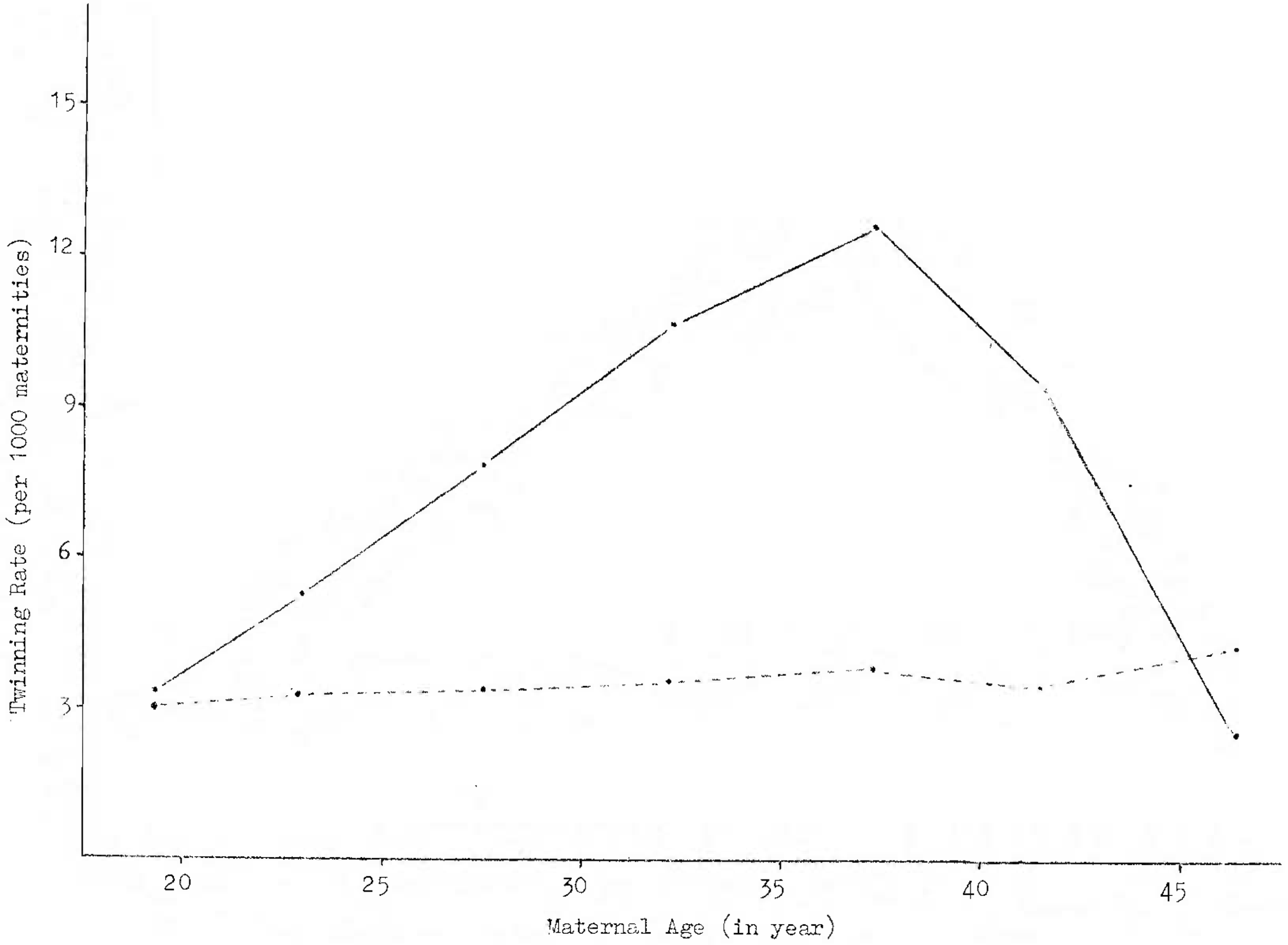

FIG. 3.2.1 Twinning and its relationship with maternal age

---.--- Twinning rate for MZ twins, ——.—— Twinning rate for DZ twins in Fig. 3.2.1, if it is descriptive of an underlying physiological effect, calls for age standardization in comparing rates. Thus, as Waterhouse writes, "that a physiological relationship exists between twin production - chiefly that of DZ twinning - and maternal age".

As far as the effect of birth rank is concerned it is seen (Waterhouse, 1950) that in general there is a steady rise of the incidence of twins with the birth rank and thus one gets only a vertical displacement between the curves for different age groups, each curve being of approximately the same shape as that of Fig. 3.2.1. For our purpose, here, we shall not complicate the issue by bringing this into the picture. But models can be separately constructed for each birth rank to take these types of variations into account.

3.3 HEREDITY AND TWINNING

Investigations on the familial incidence of twinning have revealed rather a conflict of evidence. Thus Weinberg (1901, 1909) concluded that dizygotic, but not monozygotic twinning was hereditary, and that the heredity involved was limited to the mother's side; on the other hand, Greulich (1935) found the hereditary influence to be at least as pronounced on the father's as on the mother's side.

Trying to resolve the problem, Bulmer (1960) concluded that "mothers, but not fathers, of twins are more often themselves twins than expected". More specifically White and Wyshak (1964) concluded, from the genealogic records at the Genealogical Society of the Church of Jesus Christ of Latter Day Saints at Salt Lake City, Utah, that when women who are DZ twins become parents they produce twins at the rate of 17.1 sets per 1000 maternities. On the other hand, the wives of men who are dizygous twins have a twinning rate of only 7.9 per 1000 maternities. This finding is consistent with that of Waterhouse (1950).

Taking all these facts into account we make the following observation to build up the models: the release of extra ova is a hereditary property i.e., mothers who are themselves members of DZ twins have more tendency to release extra ova, in the early stage of the period of gestation, than mothers who are coming out from a single ovum (single birth or MZ twins). Bulmer (1958), Das (1953-56) or Allen (1960) took no account of this fact while making attempts to lay down their models for predicting the frequencies of multiple birth.

3.4 A PROBABILISTIC MODEL FOR MULTIPLE BIRTHS

In this section a model incorporating the heredity of twinning has been presented and methods of estimating the underlying parameters are indicated. This also takes care of the criticism of Das (1956) that the twinning scission of a zygote is taken as a constant in the literature. In the sequel we call a female to be of type A if during her birth only one ovum is released (i.e., either she comes as a result of a single birth or she is a member of single zygous multiple birth). Otherwise she is said to be of type B (i.e., she is a member of a multiple birth resulting from more than one ovum).

We now make the following assumptions:

(1) The chance of an extra ovulation, for a type A mother, in a short time interval is constant and independent of the number of extra eggs which have already been released; an ovary producing extra eggs is supposed to be like a radio-active substance emitting α -particles. Now it is well known (see for example Feller, 1950) that such a process obeys a Poisson distribution

$$\frac{e^{-\mu} \mu^{r}}{r!}$$

where μ is the mean of the distribution and r is the number of events. In the present case the parameter is taken to be $x_1 (x_1 > 0)$, i.e., Prob. (i extra ovulation takes place | mother is of type A) = $x_1^{\,i} e^{-x} / i!$ ... (3.4.1) where, x_1 is the intensity of extra ovulation for a type A mother.

(2) Same argument leads us to assume that the release of extra ova from a type B mother also follows a Poisson distribution with parameter $x_2 (x_2 > x_1 > 0)$ i.e., Prob. (i extra ovulation takes place | mother is of type B) = $x_2^{\,i} e^{-x_2} / i!$ ... (3.4.2) where, x_2 is the intensity of extra ovulation for a type B mother.

(3) The chance that a particular embryo divides during a short time interval is also supposed to be constant and independent of the number of divisions which have already occurred; but in computing the chance that some embryo divide during a short time interval, this figure must be multiplied by the number of embryos at risk. This is a realization of the pure birth process known as the Yule Process, first studied by Yule (1924) in connection with the mathematical theory of evolution. Thus, if such a process starts with i zygotes at zero time, then the chance that it has grown by successive division to n zygotes at unit time is

$$\binom{n-1}{n-i} e^{-ip} (1 - e^{-p})^{n-i} \quad \text{for} \quad n \geq i \quad \text{and } i \geq 1 \qquad (3.4.3)$$

where, p is the intensity of scission for any particular zygote ($p > 0$). We further assume that this parameter p does not depend upon the type of the mother.

(4) The release of an extra egg and the splitting of an embryo are two independent random events.

Under these assumptions :

Prob. (twin birth) = Prob. (twin birth | mother is of type A). Prob. (mother is of type A) + Prob. (twin birth | mother is of type B) . Prob. (mother is of type B).

Now let P be the probability that the mother is of type A.

Hence P = Prob. (mother is of type A | Grandmother is of type A).
Prob. (Grandmother is of type A) + Prob. (mother is of type A | Grandmother is of type B). Prob. (Grandmother is of type B)
= Prob. (during mother's borth only one ovum is released and the possible multiple birth is by scission of a single resulting zygote | Grandmother is of type A).
P + Prob. (during mother's birth ...... of a single resulting zygote | Grandmother is of type B). $(1 - P)$.

/Assuming that the probability that a female is of type A does not change over the generation].

Hence, $P = e^{-x_1} . P + e^{-x_2} (1 - P)$

i.e., $P = e^{-x_2} / (1 - e^{-x_1} + e^{-x_2})$ ... (3.4.4)

Now, Prob. (twin birth | mother is of type A)

= Prob. (one ovum is released and the resulting zygote splits just once | mother is of type A) + Prob. (only one extra ovulation takes place and there is no scission of the resulting pair of zygotes | mother is of type A)

$$= e^{-x_1-p} (1 - e^{-p}) + x_1 e^{-x_1} e^{-2p} \qquad (3.4.5)$$

and by similar argument,

Prob. (twin birth | mother is of type B)

$$= e^{-x_2-p} (1 - e^{-p}) + x_2 e^{-x_2} e^{-2p} \qquad (3.4.6)$$

Thus, Prob. (twin birth)

$$= [e^{-x_1-p} (1 - e^{-p}) + x_1 e^{-x_1 - 2p}].P$$
$$+ [e^{-x_2-p} (1 - e^{-p}) + x_2 e^{-x_2 - 2p}] (1 - P)$$

from (3.4.5) and (3.4.6)

$$= \frac{e^{-x_2 - p}}{1 - e^{-x_1} + e^{-x_2}} [x_2 e^{-p} + (x_1 - x_2) e^{-x_1-p} + (1 - e^{-p})]$$

using (3.4.4). (3.4.7)

A close look at (3.4.7) will automatically reveal that

$$\text{Prob. (MZ twin birth)} = \frac{e^{-x_2 \; -p}(1 - e^{-p})}{1 - e^{-x_1} + e^{-x_2}}$$

and

$$\text{Prob. (DZ twin birth)} = \frac{e^{-x_2 \; - 2p}\left[x_2 + (x_1 - x_2)\, e^{-x_1}\right]}{1 - e^{-x_1} + e^{-x_2}}$$

Hence, among the twin births MZ and DZ pairs bear a ratio of

$$e^{x_2}\,(e^{p} - 1) : (x_1 - x_2 + x_2\, e^{x_1}) \qquad (3.4.8)$$

Remark 3.4.1 From (3.4.7), by taking

$$x_1 = x_2 = x \quad \text{and} \quad e^{-p} \doteq 1, \; e^{-x_1} \doteq 1, \; 1 - e^{-p} = Y$$

as was done by Bulmer (1958) we obtain Bulmer's formula for twin birth where p and x were virtually interpreted as the probabilities for MZ and DZ twin births.

The probability of r-tuple births can also be easily deduced from the above considerations as follows:

Prob. (r - tuplebirths)

$= \sum_{i=1}^{r}$ Prob. (r zygotes are formed at the end $|$ i zygotes are present initially). Prob. (i zygotes are present initially).

But, Prob. (i zygotes are present initially)

= Prob. (i-1 extra ovulation takes place | mother is of type A). Prob. (mother is of type A) + Prob. (i-1 extra ovulation takes place | mother is of type B). Prob. (mother is of type B).

$$= \frac{x_1^{i-1} e^{-x_1}}{(i-1)!} \cdot P + \frac{x_2^{i-1} e^{-x_2}}{(i-1)!} (1-P).$$

Hence, Prob. (r - tuple births)

$$= \sum_{i=1}^{r} \binom{r-1}{r-i} e^{-ip} (1 - e^{-p})^{r-i} \left[\frac{e^{x_1} x_1^{i-1}}{(i-1)!} P + \frac{e^{-x_2} x_2^{i-1}}{(i-1)!} (1-P) \right] \tag{3.4.9}$$

Remark 3.4.2 From (3.4.9) also by similar approximations as earlier, we get Bulmer's formula as a special case. However, these approximations do not appear to be satisfactory from a mathematical point of view.

Remark 3.4.3 It might be noted that each term in the summation of (3.4.9) is the probability of a i-zygous r-tuple birth. Thus, it takes care of the criticism of Das (1953-55) and Allen (1960) and provides a formula involving lesser number of parameters than used by Allen (1960).

Remark 3.4.4 The hereditary property of the release of extra eggs indicated in section 3.3 translates itself as $x_1 < x_2$ in the assumption 2 of this section. It is to be noted that mothers who are members of a single-zygous multiple birth are not differentiated from those who come from single births.

Estimation of the parameters :

In order to estimate the parameters x_1, x_2 and p let us consider the following model: since twins are more commonly observed than any other higher multiple birth, we shall present a method of estimating the parameters from a random sample of twins. Suppose r is the sex-ratio in the twins (i.e., the probability of a male). There are some controversies about the choice of sex-ratio (Das, 1953) arising out of the differential prenatal mortality for male (XY) and female (XX) zygotes and in what follows r is considered as the secondary sex-ratio i.e., the sex-ratio at birth.

Thus, Prob. (male - male twins) = r^2

Prob. (male - female twins) = $2r(1 - r)$

and Prob. (female - female twins) = $(1 - r)^2$

in case of DZ twins and Prob. (male - male twins) = r and Prob. (female - female twins) = $(1 - r)$ in case of MZ twins.

Also let

ϕ_1 = Prob. (MZ twins | mother is of type A) = $e^{-x_1 - p}(1 - e^{-p})$

ϕ_2 = Prob. (DZ twins | mother is of type A) = $x_1 e^{-x_1 - 2p}$

Ψ_1 = Prob. (MZ twins | mother is of type B) = $e^{-x_2 - p}(1 - e^{-p})$

Ψ_2 = Prob. (DZ twins | mother is of type B) = $x_2 e^{-x_2 - 2p}$

Now, a twin can belong to one of the ten mutually exclusive and collectively exhaustive classes whose probabilities can be expressed in terms of ϕ_1, ϕ_2, Ψ_1, Ψ_2 and r. Table 3.4.1 gives the different classes and the class probabilities.

Note that R = Prob. (twin birth)

$$= P(\phi_1 + \phi_2) + (1 - P)(\Psi_1 + \Psi_2)$$

Thus one has a classical multinomial distribution with ten classes and involving the unknown parameters x_1, x_2, p and r. Before going through the actual estimation of the parameters let us present some obvious relationships among the ϕ 's and Ψ's.

TABLE 3.4.1

Twin births and their probabilities

Mother's type	Twin births				
	Monozygotic		Dizygotic		
	MM	FF	MM	MF	FF
A	$\frac{P\phi_1 r}{R}$	$\frac{P\phi_1(1-r)}{R}$	$\frac{P\phi_2 r^2}{R}$	$\frac{2P\phi_2 r(1-r)}{R}$	$\frac{P\phi_2(1-r)^2}{R}$
B	$\frac{(1-P)\Psi_1 r}{R}$	$\frac{(1-P)\Psi_1(1-r)}{R}$	$\frac{(1-P)\Psi_2 r^2}{R}$	$\frac{2(1-P)\Psi_2 r(1-r)}{R}$	$\frac{(1-P)\Psi_2(1-r)^2}{R}$

M = Male ; F = Female

From the definition of ϕ_1, ϕ_2, Ψ_1 and Ψ_2, it is easy to see that

$$\frac{x_2}{x_1} = \frac{\phi_1 / \phi_2}{\Psi_1 / \Psi_2} \tag{3.4.10}$$

$$p = \log \left[1 + x_1 \cdot \frac{\phi_1}{\phi_2} \right] \tag{3.4.11}$$

and

$$\frac{1 - P}{P} = \frac{1 - e^{-x_1}}{e^{-x_2}} \tag{3.4.12}$$

Let $n_1, n_2, \ldots, n_{10}$ be the observed frequencies of the ten different types of twins from a random sample of N twins (see Table 3.4.2). Then, one can easily see that an estimate of r is given by

$$\hat{r} = \frac{n_1 + 2n_3 + n_4 + n_6 + 2n_8 + n_9}{N + (n_3 + n_4 + n_5) + (n_8 + n_9 + n_{10})} \qquad (3.4.13)$$

It is worthwhile to note that this is also the maximum likelihood estimate of r. The other likelihood equations lead us to

$$\frac{\Phi_1}{\Phi_2} = \frac{n_1 + n_2}{n_3 + n_4 + n_5}$$

and

$$\frac{\psi_1}{\psi_2} = \frac{n_6 + n_7}{n_8 + n_9 + n_{10}}$$

and hence, using (3.4.10), we have

$$\frac{x_2}{x_1} = \frac{n_1 + n_2}{n_3 + n_4 + n_5} \cdot \frac{n_8 + n_9 + n_{10}}{n_6 + n_7} \qquad (3.4.14)$$

From the likelihood equations, one can also get

$$\frac{n_1 + n_2}{n_6 + n_7} = \frac{1}{e^{x_1} - 1} \qquad \text{(using (3.4.12)}$$

$$\text{Hence, } \hat{x}_1 = \log \left[1 + \frac{n_6 + n_7}{n_1 + n_2} \right] \qquad (3.4.15)$$

and hence $\hat{x}_2$ can be obtained from (3.4.14).

(3.4.11) enables us to get the estimate of p as

$$\hat{p} = \log \left[1 + \hat{x}_1 \left(\frac{n_1 + n_2}{n_3 + n_4 + n_5} \right) \right] \qquad (3.4.16)$$

We now proceed to illustrate the estimation procedure through an hypothetical data shown by Table 3.4.2. This table also shows the expected frequencies of the different classes as obtained from the estimated parameters.

TABLE 3.4.2

Observed and expected frequencies of twin briths

Mother's Type		Twin births				
		Monozygotic		Dizygotic		
		MM	FF	MM	MF	FF
A	Obs.	1439(n_1)	1408(n_2)	1803(n_3)	3547(n_4)	1745(n_5)
	Exp.	1435.83	1411.17	1804.61	3547.23	1743.15
B	Obs.	4(n_6)	4(n_7)	20(n_8)	39(n_9)	19(n_{10})
	Exp.	4.03	3.97	19.84	39.00	19.16

Using the equations (3.4.13) to (3.4.16) we obtain

$$\left.\begin{aligned} \hat{r} &= 0.5043 \\ \hat{x}_1 &= 0.0028 \\ \hat{x}_2 &= 0.0110 \\ \text{and} \qquad \hat{p} &= 0.0011 \end{aligned}\right\} \qquad (3.4.17)$$

Since these estimates are almost efficient (as they are, in a sense, derived from maximal likelihood equations) one can proceed to see the goodness of fit. The value of χ^2 with 5 degrees of freedom turns out to be 0.021 ($P > 0.75$).

The estimate of the propertion of MZ twins out of all twin birth is (using formula (3.4.8)) 0.2870 which is fairly close to the observed value of 2855/10028 = 0.2847.

3.5 A STOCHASTIC MODEL FOR MULTIPLE BIRTHS

In this section we shall assume that the probabilities of the release of extra eggs as well as the scission of the zygotes are functions of the time after commencement of the period of gestation. Thus, the above probabilities depend on the period of conception. Let $\mu(t)$ denote the intensity with which a scission takes place and $\lambda(t)$, the intensity of extra ovulation at time t

after the commencement of gestation. Suppose we also consider the prenatal mortality whose intensity is $D(t)$ at time t. Needless to say that the processess like scission of a zygote and extra ovulation stop after a certain period (say, t_1) whereas the prenatal mortality continues to be in operation throughout the period of conception.

Thus we have the following probabilities :

(i) $i + 1$ zygotes at time $t + \delta t$ starting with i zygotes at time t. - - - - - $[\lambda(t) + \mu(t)]. \delta t$

(ii) $i - 1$ zygotes at time $t + \delta t$ starting with i zygotes at time t. - - - - - $D(t). \delta t$

(iii) i zygotes at time $t + \delta t$ starting with i zygotes at time t. - - - - - - - $[1 - \lambda(t) - \mu(t) - D(t)] \delta t$

(iv) $i \pm r$ zygotes at time $t + \delta t$ starting with i zygotes at time t. - - - - $0(\delta t)$ for $r \geq 2$.

Also suppose that the process starts with a single zygote i.e., Prob. (one zygote is formed initially) $= P_1(0) = 1$ and $P_r(0) = 0$ for $r \neq 1$.

Then from the classical procedure, we obtain the differential equation

$$\frac{d\,P_i(t)}{dt} = P_{i-1}(t).(i - 1).B(t) + P_{i+1}(t).(i + 1).D(t)$$

$$- P_i(t) \,.\, i \,.\, [B(t) + D(t)] \qquad (3.5.1)$$

and $\dfrac{d\,P_0(t)}{dt} = D(t) \,.\, P_1(t)$ for all $i \geq 1$

where $B(t) = \lambda(t) + \mu(t)$.

A solution of equation (3.5.1) is given by

$$\left.\begin{array}{l} P_i(t) = [1 - \alpha(t)].[1 - \beta(t)].[\beta(t)]^{i-1}, \; i \geq 1 \\ P_0(t) = \alpha(t) \end{array}\right\} \qquad (3.5.2)$$

where $\alpha(t) = 1 - \dfrac{e^{-\gamma(t)}}{w(t)}$

$$\beta(t) = 1 - \frac{1}{w(t)}$$

$$\gamma(t) = \int_0^t [D(\tau) - B(\tau)]\, d\tau$$

and $w(t) = e^{-\gamma(t)} [1 + \int_0^t D(\tau)\, e^{\gamma(\tau)}\, d\tau]$

(Kendall (1948)).

In the present case, we assume that the intensity of extra ovulation is the same for all mothers and is given by

$$\lambda(t) = \text{max.} \left(0, \lambda \cdot \frac{t_1 - t}{t_1}\right)$$

where, $\lambda > 0$ is a constant and t_1 is the period upto which extra ovulation or scission is possible. Let the intensity of scission be given by

$\mu(t) = \text{max.} \left(0, \frac{\mu}{1 + \mu . t} \cdot \frac{t_1 - t}{t_1}\right)$ where μ is a constant. This gives the rate of scission per zygote at time t. The factor $\mu/(1 + \mu t)$ can be considered as analogous to the decay factor in Polya Process.

Let the intensity of prenatal mortality be υ which is a positive constant.

Remark 3.5.1 It may be observed that both $\lambda(t)$ and $\mu(t)$ are decreasing functions of t.

Remark 3.5.2 Since the primary interest in this study is the number of live births, there is no loss of generality in assuming that the intensity of prenatal mortality is a constant. However it must be borne in mind that this assumption cannot be extended further since postnatal mortality rates are vastly different from prenatal mortality rates.

<u>Remark 3.5.3</u> It is clear that had we not considered the prenatal mortality, the resulting process would be the generalized Poisson process as follows: Suppose the intensity of increase of zygotes (either by scission or by release of extra eggs) be $B(t)$. Then with the same initial conditions as above, one easily obtains that

$$P_i(t) = e^{-K(t)} \cdot \frac{[K(t)]^i}{i\,!}$$

where $K(t) = \int_0^t B(\tau)\, d\tau$.

In the above case of generalized birth and death process $B(t) = \lambda(t) + \mu(t)$. For any time $t < t_1$

$$\gamma(t) = \int_0^t [D(\tau) - B(\tau)]\, d\tau$$

$$= \frac{1}{t_1} \Big[((\upsilon - \lambda)\, t_1 + 1)t + \frac{\lambda t^2}{2} - \Big(\frac{1 + \mu t_1}{\mu}\Big) \log(1 + \mu t) \Big]$$

$$\int_0^t e^{\gamma(\tau)} \, d\tau$$

$$= e^{-\frac{1}{2}u} \left[\int_0^t e^{\frac{\lambda}{2t_1}(\tau + u)^2} (1 + \mu\tau)^{-(1 + \frac{1}{\mu t_1})} \, d\tau \right]$$

where $u = \frac{1}{\lambda} [t_1 (\upsilon - \lambda) + 1]$.

A closed expression for this integral on the right hand side is not available even though it is clear that the finite integral exists (since the integral is a bounded continuous function in the range $(0, t)$). This may be evaluated by some approximation formula. We denote

$$\int_0^t \upsilon \, e^{\gamma(\tau)} \, d\tau = u(t).$$

Thus $w(t) = e^{-\gamma(t)} [1 + u(t)]$

$$\alpha(t) = u(t) / [1 + u(t)]$$

$$\beta(t) = 1 - e^{\gamma(t)} / [1 + u(t)]$$

and hence by substitution in (3.5.2) one gets

$$\left.\begin{array}{l} P_i(t) = e^{\gamma(t)} \left[1 - \frac{e^{\gamma(t)}}{1 + u(t)} \right]^{i-1} . \, (1 + u(t))^{-2} ; \; i \geq 1 \\ \text{and} \quad P_0(t) = u(t) / [1 + u(t)] \quad \text{for} \quad t \leq t_1 \end{array}\right\} \quad (3.5.3)$$

Now suppose $t > t_1$

$$\gamma\,(t) = \gamma\,(t_1) + \upsilon\ .\ (t - t_1)$$

$$w\,(t) = e^{-\gamma(t)} \left[1 + u(t_1) + e^{\gamma(t)} - e^{\gamma(t_1)} \right]$$

$$\alpha\,(t) = 1 - \left[\frac{e^{-\gamma(t)}}{e^{-\gamma(t)} \left[1 + u(t_1) + e^{\gamma(t)} - e^{\gamma(t_1)} \right]} \right]$$

$$= \frac{u(t_1) + e^{\gamma(t)} - e^{\gamma(t_1)}}{1 + u(t_1) + e^{\gamma(t)} - e^{\gamma(t_1)}}$$

and

$$\beta\,(t) = \frac{1 + u(t_1) - e^{\gamma(t_1)}}{1 + u(t_1) + e^{\gamma(t)} - e^{\gamma(t_1)}}$$

Thus again by substitution in (3.5.2)

$$\left. \begin{aligned} P_i(t) &= \frac{e^{\gamma(t)} \left[1 + u(t_1) - e^{\gamma(t_1)} \right]^{i-1}}{\left[1 + u(t_1) + e^{\gamma(t)} - e^{\gamma(t_1)} \right]^{i+1}} \quad \text{for } i \geq 1 \\ \text{and} \quad P_0(t) &= \frac{u(t_1) + e^{\gamma(t)} - e^{\gamma(t_1)}}{1 + u(t_1) + e^{\gamma(t)} - e^{\gamma(t_1)}} \end{aligned} \right\} (3.5.4)$$

If one wants to incorporate the differential intensities of release of extra eggs for type A and type B mothers as was done in the earlier model this can be easily done as follows :

$$P_i(t) = P_i\,(t \mid \text{mother is of type A}).\ \text{Prob. (mother is of type A)} + P_i\,(t \mid \text{mother is of type B}).\ \text{Prob. (mother is of type B)}. \tag{3.5.5}$$

Now, $P_i(t \mid$ mother is of type A) can be obtained by inserting the corresponding intensity of extra ovulation, say $\lambda_1(t)$, and similarly $P_i\,(t \mid$ mother is of type B) can also be obtained. However, this involves another unknown parameter P = Prob. (mother is of type A). Note that the probability of any specified i-zygous r-tuplet birth cannot be derived from the above model unlike in the probabilistic model considered in section 3.4. Thus, in particular, P cannot be derived from the model and this has to be estimated from the sample.

Estimation of parameters

From the foregoing discussion it is evident that the ultimate model (3.5.4) involves two unknown parameters $\gamma(t)$ and υ . Similarly the model (3.5.5) involves four unknown parameters. These parameters can be estimated by the standard method of maximum likelihood estimation under the following assummtion : We tacitly assume that the period of conception is the same for all mother, say equal to t_0. Then $P_i(t_0)$ will give the probability of i-tuple births for all i. However, since 0 births cannot be observed, we shall have a truncated distribution.

$$\text{Prob. (i-tuple births)} = \frac{P_i(t_0)}{1 - P_0(t_0)} \quad \text{for} \quad i \geq 1.$$

$$= \frac{e^{\gamma(t_0)} \cdot [1 + u(t_1) - e^{\gamma(t_1)}]^{i-1}}{[1 + u(t_1) + e^{\gamma(t_0)} - e^{\gamma(t_1)}]^{i}} \qquad (3.5.6)$$

for $i \geq 1$.

Now consider a random sample of n births and the likelihood function of the sample can be written as usual.

However, we present here a quick method of estimating the parameters and hence predicting the chance of higher multiple births using some approximate formulae.

Neglecting second and higher order terms in υ , λ or μ , we have

$$e^{\gamma(t_1)} \doteq 1 + (\upsilon - \frac{\lambda}{2} - \mu)\, t_1$$

$$e^{\gamma(t_0)} \doteq 1 + \upsilon\, t_0 - (\lambda/_2 + \mu)\, t_1 .$$

and $$u(t_1) \doteq \upsilon\, t_1 .$$

Hence, from (3.5.6) we have

$$\text{Prob. (i-tuple birth)} \doteq (1 - \alpha)\ \alpha^{i-1} \qquad i \geq 1$$

$$\text{where,}\quad \alpha = \frac{(\frac{\lambda}{2} + \mu)\, t_1}{1 + \mu t_0}$$

An estimate of α (by method of moments) is given by

$$\hat{\alpha} = \frac{m - 1}{m} \tag{3.5.7}$$

where the norm of the nature of births in the sample is a m-tuple.

Table 3.5.1 represents the data, published in Bulmer (1958), of multiple births in England and Wales during 1938 to 1955. The value of m, as computed from the data, is 1.0124. Hence

$$\hat{\alpha} = 0.0123 \text{ (using 3.5.7)}.$$

Table 3.5.1 also represents the predicted number of different types of births as reported by Bulmer as well as obtained through the model described in this section.

TABLE 3.5.1

Number of different types of births as observed, and as calculated by different authors, for England and Wales

Types of births	Frequency		
	Observed	Predicted by Bulmer	Predicted by the present model
Single	12,149,571	12,147,573	12,150,061
Twins	150,072	151,946	149,076
Triplets	1,336	1,460	1,341
Quadruplets	21	21	22
Total	12,301,000	12,301,000	12,301,000

From the table, it can be seen that the present model predicts the frequencies of the different types of births more accurately than Bulmer's one. The same statement is found to be true when mother's of different age-groups were considered separately. A comparison with other models (e.g., Jenkins, 1927; Das, 1953-56 or Allen, 1960) are not presented here, as it was seen that they deviate from the observed frequencies by greater margins.

3.6 AGE-DEPENDENT MODEL FOR MULTIPLE BIRTHS

In the model, discussed in section 3.5, it has been assumed implicitly that the intensities of extra ovulation is the same for mothers of all age. However, taking the evidences of section 3.2 into account, it is observed that these intensities vary with the age of the mother. The tendency of extra ovulation increases with the age and then again decreases whereas the tendency of scission of zygotes slowly increases (which we may, for all practical purposes, take to be a constant) with the age of the mother. It is also to be observed that these facts are true for mothers of all types. In this section a stochastic model which takes this fact into account is constructed. For this purpose, we shall not make a distinction in the type of mother (for simplicity).

Thus we have the starting point as follows: For any mother of age 's', (i) let $\lambda(s, t)$ be the intensity of release of extra eggs at time t - after the commencement of conception, where for each fixed 't' $\lambda(s, t)$ satisfies the above requirement and for each fixed 's' $\lambda(s, t)$ is as in section 3.5. (ii) since the intensity of scission does not change with the age of the mother, as in section 3.5 one can assume that $\mu(t)$ as defined there, is the intensity of scission. For similar reasons let $D(t)$ denote the intensity of the prenatal mortality.

Then the probability $P_i(t)$ of i-tuple births is obtained exactly as in section 3.5, expression (3.5.4), with $\lambda(s, t)$ substituted for $\lambda(t)$.

One can try many suitable forms of $\lambda(s, t)$ as a function of s for all t. From the empirical studies, as indicated in section 3.2, one can assume, for example, the form

$$\lambda(s, t) = \text{Max.}\ (0,\ \lambda\, e^{-\alpha (s_1 - s)} \cdot \frac{t_1 - t}{t_1})\quad \text{for}\quad s \leq s_1$$

$$= \text{Max.}\ (0,\ \lambda\, e^{-\alpha s_1} \cdot \frac{t_1 - t}{t_1})\ \text{for}\quad s > s_1$$

where α , λ are positive constants and s_1 is the average age of the mother where the intensity of release of extra eggs is maximum. Bulmer as well as Stocks observe that the process starts at the age of about 19 years and assumes its maximum at the age of 37 years (s_1) and then there is a sudden fall till the age of menopause.

Remark 3.6.1 A model considering the differential intensities of extra ovulation for different types of mothers can be obtained exactly as in section 3.5 and hence details will not be presented here.

Remark 3.6.2 Through the form of $\lambda(s, t)$, as suggested above is discontinuous at $s = s_1$, but it does not bring into any complication, since for mothers whose age, $s \leq s_1$ the process is automatically different from the one where mother's age, $s > s_1$. However, one can do away with this difficulty if, instead of bringing the exponential term into consideration, he fits a suitable polynomial form in s.

For example, in case of the data for England and Wales (Stocks, 1952; see Table 3.2.1) one can use form of $\lambda(s, t)$ given by :

$$\lambda(s, t) = \text{Max.}\ (0, \lambda \cdot \frac{t_1 - t}{t}(a_0 + a_1 s + a_2 s^2))$$

where $a_0 = -41.0269$

$a_1 = 3.1364$

and $a_2 = -0.0468$.

The parameters, in such a model, can also be estimated in a similar manner. However, the method of maximum likelihood turns out to be laborious in this section and method of iteration has to be used, though it pays the dividend that these estimators can be used for further testing purposes like test of goodness of fit. If one is interested only in getting the estimates of the parameters, other methods like the method of moments may reduce the labour. It leaves therefore, a scope of further analytical as well as empirical studies in this direction.

CHAPTER IV

GENE FREQUENCY ESTIMATES IN BLOOD GROUPS

4.0 INTRODUCTION

Estimation of blood group gene frequencies was the immediate problem after the discovery of the different blood group systems. This problem of computing the relative frequencies of the genes governing the blood groups in a population had great contributions from F. Bernstein, A. S. Wiener, R. A. Fisher, C. R. Rao, W. C. Boyd apart from many others. Because of the volume of the contributions it is quite difficult to have a review of the works in a short space. Therefore in this chapter we shall consider only two blood group systems, namely, ABO and MNSs system and study some of the statistical properties of the estimators arising out of the different estimation procedures. For MNSs system we shall present maximum likelihood estimation procedure using two generation data, the same for A-B-O system has already been discussed in Chapter II of the thesis. It may be noted that though our discussions in sections 4.1.1 to 4.1.4 are always with reference to the genetics of A-B-O blood group system, results obtained herein are also valid for any character which has an equivalent phenotype-genotype relationship. This follows from Cotterman (1953) who coined the term 'phenogram' to formulate genotype-phenotype relations as simply as possible, especially when multiple alleles are involved. Thus the phenogram 3-4-1 implies a genetic system consisting of 3 alleles, 4 phenotypes and a serial number 1. The last number is arbitrary, but is necessary due to the fact that there is more than one phenotype system having specified numbers of alleles and phenotypes. Under this terminology, the standard ABO blood-group system in man has the phenogram 3-4-1 and possesses the same statistical properties as the leaf pigmentation system in Coleus, which consists of three alleles P, p^G and p, and which generates four phenotypes: green (p^Gp^G and p^Gp), purple (PP and Pp), grey (Pp^G) and pattern (pp) (Boye, 1941). An equivalent system is also found in butterfly Neozyphyrus taxila with respect to four types of the forewing in females (Komai, 1953). To this effect, our discussions on ABO blood-group system is also valid for any character having a phenogram of the type 3-4-1.

4.1 ABO BLOOD GROUP SYSTEM

4.1.0 Genetics of ABO blood groups :

The existence of four main groups of human beings have been well established with respect to ABO 'isoagglutination'. The term agglutination refers to the clumping of the blood cells, and the prefix iso (from the Greek 'isos', meaning equal) signifies that agglutination is caused by sera from the same species, man.

These four groups of persons are differentiated from one another by the immunological properties of both their red cells and their serum. The red cells of an individual possesses either one or the other, both, or neither of two substances (or groups of substances) called 'antigens', or 'agglutinogens', A and B; and his serum possesses either one or the other, neither, or both of two substances (or groups of substances) called antibodies, or agglutinins, anti-B and anti-A. Red cells containing antigen A are agglutinated by anti-A, cells containing antigen B by anti-B. The four groups of persons are named after their antigens: A, B, AB, and O. Not only does every kind of human blood lack the antibodies which would agglutinate its own red cells - a necessary condition, since any appreciable clumping would be fatal - but also every kind of human blood contains these antibodies which are compatible with the antigens of its cells. This is depicted in Table 4.1.

TABLE 4.1

ABO blood groups with the antigens and antibodies present in them

Blood Group	Antigens present on red cells	Antibodies present in plasma (or serum)
O	None	Anti-A, Anti-B
A	A	Anti-B
B	B	Anti-A
AB	A, B	None

It is obvious that the existence of four blood groups means that the gene that controls them must have more than two allelic forms, since two alleles may, at most, give rise to three different phenotypes. The now well-established explanation that three multiple alleles govern the inheritance of the main blood groups was historically preceded by another hypothesis. This was based on the assumption that an individual's blood group was determined by two genes at independent loci in two pairs of chromosomes. However, contradictions to this theory were not apparent for many years and in fact the replacement of this hypothesis by the theory of multiple alleles was based on a consideration of the relative frequencies of the four types of individuals in various population. This was due to the mathematician Felix Bernstein (1878-1956). A full account to the two-gene-pair hypothesis and its contradictions are presented in Stern (1960, pp. 179-182).

Having proven the inadequacy of the two-gene hypothesis, Bernstein assumed the existence of three multiple alleles, called I^A, I^B, and I^O (we shall, here, use the symbols as A, B and O). It was further assumed that the alleles A and B are codominant if combined in the genotype AB, but that either allele is dominant in heterozygous combination with O.

From these assumptions, the frequencies of the six possible genotypes can be written down under Model I and Model II and are presented in Table 4.2.

TABLE 4.2

ABO blood group phenotypes, genotypes and their relative frequencies

Phenotype	Genotype	Relative Frequency	
		Model I	Model II
O	OO	r^2	$r^2 + Fr(1 - r)$
A	OA	$2pr$	$2(1 - F)pr$
	AA	p^2	$p^2 + Fp(1 - p)$
B	OB	$2qr$	$2(1 - F)qr$
	BB	q^2	$q^2 + Fq(1 - q)$
AB	AB	$2pq$	$2(1 - F)pq$

Note : p, q and r are the frequencies of the allelomorphs A, B and O respectively and F is a constant (0 F 1) with the same interpretations as in the earlier cases.

4.1.1 Estimation of gene frequencies assuming model I population structure :

The first attempt to obtain the gene frequencies for ABO blood group system was made by Bernstein (1925) who suggested that the estimates of p, q and r could be derived in the following manner : We can estimate the frequency of the gene, O, directly from the proportion of individuals of phenotype O, thus if $\overline{O}$, $\overline{A}$, $\overline{B}$ and $\overline{AB}$ represent the observed frequencies of the four phenotypes adding to N, r^2 is given by

$$r^2 = \frac{\overline{O}}{N}$$

and hence $r' = \sqrt{\frac{\overline{O}}{N}}$ (4.1.1)

The proportion of type A individuals is

$$\frac{\overline{A}}{N} = p^2 + 2pr ;$$

and, if we add the proportion of type O individuals to both sides of this equation, then

$$\frac{\bar{A} + \bar{O}}{N} = p^2 + 2pr + r^2$$

and $$p + r = \sqrt{\frac{\bar{A} + \bar{O}}{N}}$$

But, since $p + q + r = 1$, we have

$$q' = 1 - \sqrt{\frac{\bar{A} + \bar{O}}{N}} \qquad (4.1.2)$$

and similarly, $$p' = 1 - \sqrt{\frac{\bar{B} + \bar{O}}{N}} \qquad (4.1.3)$$

The expressions in (4.1.1) to (4.1.3) are known as Berstein's unadjusted estimates. One may note a few statistical properties of these estimates :

(1) It is easy to see that these estimators are consistent for p, q, r in their admissible range.

(2) The sum total of p', q', r' need not be one always. In fact $p' + q' + r' = 1$ if

$$\frac{2\bar{A}}{N} + \frac{\overline{AB}}{N} = 2\left[\sqrt{\frac{\bar{A} + \bar{O}}{N}} - \sqrt{\frac{\bar{O}}{N}} \cdot \sqrt{\frac{\bar{B} + \bar{O}}{N}}\right]$$

(3) It may further be noted that these estimates are inefficient. However, this problem will be dealt with in greater details in sequel.

Bernstein proceeded a step further to solve the second type of difficulty as follows : let D be the difference between 1 and the sum of the estimates, that is

$$D = 1 - p' - q' - r'$$

$$= \sqrt{\frac{\overline{B} + \overline{O}}{N}} + \sqrt{\frac{\overline{A} + \overline{O}}{N}} - \sqrt{\frac{\overline{O}}{N}} - 1.$$

Interms of p', q' and r' we find

$$\frac{\overline{A} + \overline{AB}}{N} = 1 - \frac{\overline{B} + \overline{O}}{N}$$

$$= \left[1 - \sqrt{\frac{\overline{B} + \overline{O}}{N}}\right]\left[1 + \sqrt{\frac{\overline{B} + \overline{O}}{N}}\right]$$

$$= p'(2 - p'). \qquad (4.1.4)$$

Similarly, $\frac{\overline{B} + \overline{AB}}{N} = q'(2 - q');$ (4.1.5)

$$\frac{\overline{A}}{N} = \frac{\overline{A} + \overline{O}}{N} - \frac{\overline{O}}{N}$$

$$= \left[\sqrt{\frac{\overline{A} + \overline{O}}{N}} - \sqrt{\frac{\overline{O}}{N}}\right]\left[\sqrt{\frac{\overline{A} + \overline{O}}{N}} + \sqrt{\frac{\overline{O}}{N}}\right]$$

$$= (p' + D)(p' + 2r' + D) \qquad (4.1.6)$$

and $\frac{\overline{B}}{N} = (q' + D)(q' + 2r' + D)$ (4.1.7)

From these he argued that better estimates of p, q, r are given by

$$\left.\begin{aligned} p_* &= p'(1 + \frac{D}{2}) \\ q_* &= q'(1 + \frac{D}{2}) \\ \text{and} \qquad r_* &= (r' + \frac{D}{2})(1 + \frac{D}{2}) \end{aligned}\right\} \tag{4.1.8}$$

which total not to 1 but to $1 - \frac{D^2}{4}$. Thus if $\frac{D^2}{4}$ is small then Bernstein's adjusted estimates, given by (4.1.8), are fairly good. We shall later on see that though these are not precisely the maximum likelihood estimates, but are fully efficient, and hence one may take their variances and covariances as those of the maximum likelihood estimates.

Wiener obtained the following expressions for estimating p, q and r from ABO data on similar grounds:

$$\begin{aligned} p' &= \sqrt{\frac{\overline{O} + \overline{A}}{N}} - \sqrt{\frac{\overline{O}}{N}} \\ q' &= \sqrt{\frac{\overline{O} + \overline{B}}{N}} - \sqrt{\frac{\overline{O}}{N}} \\ \text{and} \quad r' &= \sqrt{\frac{\overline{O}}{N}} \end{aligned} \tag{4.1.9}$$

One may note that like Bernstein's unadjusted estimates, these estimates are also consistent but not efficient and by necessity they need not add to one.

An adjustment to these estimates are provided by Fisher as

$$p'' = \left[\sqrt{\frac{\bar{O}+\bar{A}}{N}} - \sqrt{\frac{\bar{O}}{N}}\right] \Big/ V$$

$$q'' = \left[\sqrt{\frac{\bar{O}+\bar{B}}{N}} - \sqrt{\frac{\bar{O}}{N}}\right] \Big/ V \qquad (4.1.10)$$

$$\text{and} \quad r'' = \left[\sqrt{\frac{\bar{O}}{N}}\right] \Big/ V$$

where, $V = \sqrt{\frac{\bar{O}+\bar{A}}{N}} + \sqrt{\frac{\bar{O}+\bar{B}}{N}} - \sqrt{\frac{\bar{O}}{N}}$

(Dobson and Ikin, 1946).

Yet another method of estimating these p, q and r ignoring the AB phenotypes can be given as follows:

$$\left.\begin{aligned} \hat{r} &= \sqrt{\frac{\bar{O}}{N}} \\ \hat{p} &= \frac{\sqrt{\frac{\bar{A}+\bar{O}}{N}} - \hat{r}\sqrt{\frac{\bar{B}+\bar{O}}{N}}}{\sqrt{\frac{\bar{A}+\bar{O}}{N}} + \sqrt{\frac{\bar{B}+\bar{O}}{N}}} \\ \text{and } \hat{q} &= \frac{\sqrt{\frac{\bar{B}+\bar{O}}{N}} - \hat{r}\sqrt{\frac{\bar{A}+\bar{O}}{N}}}{\sqrt{\frac{\bar{A}+\bar{O}}{N}} + \sqrt{\frac{\bar{B}+\bar{O}}{N}}} \end{aligned}\right\} \qquad (4.1.11)$$

Note that unlike Bernstein and Wiener estimates, these estimates always add upto 1. However, the variances of these estimates are much complicated.

The most efficient procedure of estimating the gene frequencies p, q and r using all the phenotypic frequencies is the method of maximum likelihood. But unfortunately explicit expressions of the estimates can not be obtained since the maximum likelihood equations can only be solved by iteration. For complete account of such an iterative scoring method one may refer to Rao (1965). However, Rao presented another algorithm which yields the maximum likelihood estimators, which appear to be more suitable for desk and electronic computers, and which avoids the repeated computation of the information matrix and its inverse.

Let p_0, q_0, r_0 be provisional estimates, and compute $P_0 = p_0 / r_0$ and $Q_0 = q_0/r_0$. Let us represent by P_k and Q_k the k^{th} approximation of p/r and q/r. The $(k+1)^{th}$ approximations are found from the formulas

$$\left.\begin{aligned} \frac{\bar{A} + \overline{AB}}{P_{k+1}} &= 2(\bar{O}) + \frac{\bar{A}}{2 + P_k} + \frac{2\bar{B}}{2 + Q_k} \\ \frac{\bar{B} + \overline{AB}}{Q_{k+1}} &= 2(\bar{O}) + \frac{2\bar{A}}{2 + P_{k+1}} + \frac{\bar{B}}{2 + Q_k} \end{aligned}\right\} \qquad (4.1.12)$$

The iteration may be repeated until stable values of P and

Q are obtained, from which the estimates of p,q and r are computed as

$$\hat{r} = \frac{1}{P + Q + 1}, \quad \hat{p} = \hat{r}\,P \quad \text{and} \quad \hat{q} = \hat{r}\,Q \; .$$

Since the equations in (4.1.12) follow straightway from the maximum likelihood equations, the estimates p , q , r given as above are almost fully efficient and for all practical purposes can be treated as equivalent to the maximum likelihood estimates.

More recently Yasuda and Kimura (1968) manipulated the maximum likelihood equations to give a gene count method which is also, basically, an iterative procedure. They claimed that estimators obtained in such a manner are also fully efficient.

4.1.2 Comparison of the estimation procedures :

We may note that Bernstein's and Wiener's estimates are consistent estimates of p, q and r, especially when these are adjusted in the manner as already discussed. Moreover, it must be borne in mind that in these two methods the AB-phenotypic frequencies are purposely ignored while computing the gene frequencies. The reason of this being, in the language of Fisher and Taylor (1940), "systematic errors, not all of which are yet understood, do undoubtedly affect the frequency of the rarest of the four blood groups, AB. As a further precaution, we have calculated the gene-ratios from the other three groups only, as in this way the effect of grouping errors is diminished". Taking one of the major causes of these systematic errors, we see that a given A gene produces less A antigen when combined in the genotype AB than in the genotype AO. Thus, at whatever level of discrimination one is working there will always tend to be an apparent deficiency of AB and thus some AB bloods will always be recorded as bloods of group B. Taking λ to be the probability of such a misclassification $(0 < \lambda < 1)$, the expected frequencies of $\overline{O}$, $\overline{A}$, $\overline{B}$ and $\overline{AB}$ in a random sample of N individuals can be given as :

$$E(\overline{O}) = N.r^2, \; E(\overline{A}) = N(p^2 + 2pr), \; E(\overline{B}) = N[q^2 + 2qr + 2\lambda pq]$$

$$\text{and} \quad E(\overline{AB}) = 2N(1 - \lambda) pq.$$

In such a case; it can easily be shown that, Bernstein's estimates p_*, q_*, r_* as given by (4.1.8) are no longer consistent. In expectation p_*, q_*, r_* differ from p, q and r respectively by amounts a, b and c given by

$$a = \tfrac{1}{2}\left[p(1-p) - p\sqrt{(1-p)^2 + 2\lambda pq} - 2\lambda pq \right]$$

$$b = \frac{q}{2}\left[\sqrt{(1-p)^2 + 2\lambda pq} - (1 - p) \right]$$

and $$c = \tfrac{1}{2}\left[\sqrt{(1-p)^2 + 2\lambda pq} - (1-p) \right]\left[\tfrac{1}{2}\sqrt{(1-p)^2 + 2\lambda pq} - \tfrac{1}{2}(1-p) + 1+r \right]$$

For studying the significance of such deficiency in AB-frequencies Fisher (in Dobson and Ikim, 1946) suggested a χ^2 statistic given by

$$\chi^2 = \frac{Z^2(\overline{A} + \overline{O})(\overline{B} + \overline{O})}{wx} \quad \text{with 1 d.f.}$$

where, $w = \nu^2$ [obtained from expression (4.1.10)]

$x = w - (\overline{O} + \overline{A} + \overline{B})$

and $Z = x - \overline{AB}$.

However, because of the simplicity of the Bernstein's and Wiener's method one may be more interested to evaluate the efficiencies of these methods. This problem was considered first by DeGroot (1956) though some of the points were mentioned by quite a few authors earlier (Bernstein, 1930; Stevens, 1938; Boyd, 1954, 1956 and Sukhatme, 1942). In his comparison, DeGroot considered the variances of each gene frequency separately rather than simultaneously dealing with all of them. In other words, his comparison is based on the variances of the estimates, and the covariances between the estimates were not brought into use. We now proceed to compare the efficiences of these methods using the concept of the generalized variances (determinant of the variance-covariance matrix). First of all, we record the covariance matrices of the estimators which we have studied earlier.

Covariance matrices of the estimates :

Stevens (1950) has derived the elements of the information matrix of the maximum likelihood estimators. DeGroot (1956) obtained the variance formulas from those after a great deal of simplification. The covariance matrix of the maximum likelihood estimates was obtained as

$$V_M = \frac{1}{2N}\begin{bmatrix} p(1-p) & -pq \\ -pq & q(1-q) \end{bmatrix} + \frac{1}{8N}\begin{bmatrix} p^2(1+\frac{r}{pq+r}) & pq(1-\frac{r}{pq+r}) \\ pq(1-\frac{r}{pq+r}) & q^2(1+\frac{r}{pq+r}) \end{bmatrix} \quad (4.1.13)$$

using the simplification of Li (1956). N, here, denotes the total number of observations in the sample.

The covariance matrix of Bernstein's unadjusted estimates is given by (Sukhatme, 1942).

$$V_B = \frac{1}{2N}\begin{bmatrix} p(1-p) & -pq \\ -pq & q(1-q) \end{bmatrix} + \frac{1}{4N}\begin{bmatrix} p^2 & pq(1 - \frac{r}{pq+r}) \\ pq(1 - \frac{r}{pq+r}) & q^2 \end{bmatrix} \quad (4.1.14)$$

and the corresponding matrix for Wiener's unadjusted estimates is seen to be

$$V_W = 2N\begin{bmatrix} p(1-p) & -pq \\ -pq & q(1-q) \end{bmatrix} + \frac{1}{4N}\begin{bmatrix} p^2(1+\frac{2q}{p^2+pr}) & pq(1+\frac{1}{pq+r}) \\ pq(1+\frac{1}{pq+r}) & q^2(1+\frac{2p}{q^2+qr}) \end{bmatrix} \quad (4.1.15)$$

The first components in (4.1.13), (4.1.14) and (4.1.15) are the covariance matrices that would have been obtained had there been no dominance, and thus, the exact number of A, B and O genes were known in the sample of 2N genes. Thus, these forms of the covariance matrices show the effect of dominance relationships between A, B, O genes on the methods of estimation.

The covariance matrices of the adjusted estimates of Bernstein and Wiener do not appear to have been published. However, Neel and Schull (1954) point out that these estimates, although not precisely maximum likelihood estimates, are fully efficient, and consequently one may use the covariance matrix of the maximum likelihood estimates for them.

Efficiences of the three methods :

The efficiency of an estimate is defined as the ratio of its variance to that of the maximum likelihood estimate of the parameter (Stevens 1938, Fisher 1950, Mather 1951). In this case we shall be using generalized variances of the estimators instead of the variance expressions.

To show that Bernstein's unadjusted estimates or Wiener's estimates are not fully efficient, suffice it to show that $|V_B|$ or $|V_W| \geq |V_M|$. One way of showing this (without computing the determinant directly) is to show that $V_B - V_M$ and $V_W - V_M$ are positive semi-definite matrices. Once this is shown the rest follows immediately (Rao, 1965).

It can be seen easily that

$$V_B - V_M = \frac{1}{8N} \cdot \frac{pq}{pq+r} \begin{bmatrix} p^2 & pq \\ pq & q^2 \end{bmatrix} \qquad (4.1.16)$$

$$V_W - V_M = \frac{1}{8N} \cdot \frac{pq}{pq+r} \begin{bmatrix} (2-p)^2 & pq+2r+2 \\ pq+2r+2 & (2-q)^2 \end{bmatrix} \qquad (4.1.17)$$

are positive semi-definite matrices. Incidentally, it may be pointed out here that the determinants of both the matrices (4.1.16) and (4.1.17) are equal to zero.

Similar comparison for Bernstein's and Wiener's methods is not worthwhile because

$$V_W - V_B = \frac{1}{4N} \cdot \frac{pq}{pq+r} \begin{bmatrix} 2(1-p) & 1+r \\ 1+r & 2(1-q) \end{bmatrix}$$

is indefinite.

It can be noted from the matrix $V_W - V_B$ that $V_W(p') \geq V_B(p')$ and $V_W(q') \geq V_B(q')$ for all values of p, q and r although $|V_W - V_B| \leq 0$ always. From the fact that $V_W - V_B$ is indefinite, it follows that there exists a linear combination of p' and q' such that

$$V_W(ap' + bq') < V_B(ap' + bq'), \qquad (4.1.18)$$

at least for some values of p, q and r.

Construction of such linear combinations will be only of mathematical interest but it is easy to show that condition (4.1.18) is satisfied for a and b determined by

$$(a + b) \left[a(1 - p) + b(1 - q) \right] < 0 \qquad (4.1.19)$$

Of course, for evaluating the efficiencies explicitely one needs to compute the determinants. Thus one gets

$$\frac{|V_B|}{|V_M|} = 1 + \frac{pq(pq + r - r^2)}{(pq + r)(pq + 2r + 2r^2)} \qquad (4.1.20)$$

That the second part on the right hand side of this equation is always positive reflects the fact that Bernstein's estimate is not fully efficient. One simple upper bound for this ratio happens to be

$$\frac{|V_B|}{|V_M|} < 1 + \frac{pq}{pq + r} \qquad (4.1.21)$$

It is evident that the ratio in (4.1.20) approaches the value 1 for p or q-values close to zero. So, for the populations where A or B genes (or both) are rare enough (e.g., American Indians and Early Europeans represented by their modern descendents the

Basques, Levin 1954) the dividend of the relatively involved computations of maximum likelihood estimates are not paid in turn.

Comparison of Bernstein's and Wiener's method gives the ratio

$$\frac{|V_W|}{|V_B|} = 1 + \frac{4p^2q^2 + pqr(4 - r^2) + 2r(1 - r^2)}{2(pq + r)^2 + r^2(pq + 2r)} \qquad (4.1.22)$$

Since the second term on the right hand side of this expression is positive for all values of p, q and r, it is clear that $|V_W|$ is always greater than $|V_B|$. But in this case the ratio has an upper bound of 3 unlike the case as described by DeGroot (1956) where the ratio $V_W(p')/V_B(p')$ grows large without bound. On the other hand the ratio is close to 1 when p or q (or both) is small. So, the Bernstein's or Wiener's estimates have high efficiency only when the frequencies of A or B genes (or both) are small.

As an example, for Indian populations (taking p, q and r as 0.18, 0.25 and 0.57 respectively; Mourant et al. 1958) one can see that Bernstein's method is as efficient as 98.85 percent whereas Wiener's method is nearly 56.69 percent efficient as compared to Bernstein's method.

4.1.3 Identifiability of Model I population :

So far we have discussed the procedures of estimating the gene frequencies p, q and r when the population has a Model I structure. This assumption is quite serious since any deviation from it will attach no physical meaning of the estimates. To say it more explicitely, one may always equate the observed O, A, B and AB phenotypic proportions to r^2, $p^2 + 2pr$, $q^2 + 2qr$ and $2pq$ respectively and solve for p, q, r subject to the condition $p + q + r = 1$. But these solutions need not be the actual proportions in which A, B, and O genes exist in the population unless the underlying population structure is of Model I type. Moreover, it must be stated that whenever in a population ABO blood groups phenotypes are seen in above stated proportions, one should not declare that the population is under panmixia with respect to ABO blood groups with A, B, O gene frequencies given by p, q and r respectively. In this section our contention is to prove a theorem showing that such a declaration will often be erroneous and hence should always be carefully avoided.

To do that let us recall the set-up of Wahlund's effect and see how that is extended to multiple alleles by Li (1969). We have seen earlier that in a population, consisting of k mendelian isolates each of which practices random mating with respect to a character expressed by two alleles A and a (whose frequencies in the i^{th} isolate are p_i and q_i, respectively), the frequencies of the three genotypes AA, Aa and aa are given by $p^2 + \sigma_p^2$, $2pq - 2\sigma_p^2$ and $q^2 + \sigma_p^2$ respectively, where $p = \Sigma w_i p_i$, is the average frequency of A-gene in the population, σ_p^2 = variance of the gene frequency (p) among the isolates, and w_i = relative size of the i^{th} isolate. We may note that $\sigma_p^2 = \sigma_q^2 = -\sigma_{pq}$ (since for all i = 1, ..., k; $p_i + q_i = 1$).

With three alleles A_1, A_2, A_3 (frequencies p_i, q_i and r_i respectively) the genotypic frequencies in the population can be designated as follows:

$$1 \quad \begin{array}{cccc} & A_1 & A_2 & A_3 \\ A_1 & p^2 + \sigma_1^2 & pq + \sigma_{12} & pr + \sigma_{13} \\ A_2 & pq + \sigma_{12} & q^2 + \sigma_2^2 & qr + \sigma_{23} \\ A_3 & pr + \sigma_{13} & qr + \sigma_{23} & r^2 + \sigma_{33} \end{array} \qquad (4.1.23)$$

where, $\sigma_1^2 = \Sigma w_i \, p_i^2 - p^2$, is the variance of A_1 gene frequency among the isolates and $\sigma_{12} = \Sigma p_i \, q_i \, w_i - pq$, is the covariance of the frequencies of A_1 and A_2, etc. Li (1969) has, further, shown that these variances and covariances are related on account of the restriction $p_i + q_i + r_i = 1$. In fact, all the covariances may be expressed in terms of the variances

$$2\sigma_{12} = \sigma_3^2 - \sigma_1^2 - \sigma_2^2$$
$$2\sigma_{13} = \sigma_2^2 - \sigma_1^2 - \sigma_3^2$$
$$2\sigma_{23} = \sigma_1^2 - \sigma_2^2 - \sigma_3^2$$

Moreover, for each row of (4.1.23), one has

$$\sigma_1^2 + \sigma_{12} + \sigma_{13} = 0$$
$$\sigma_{12} + \sigma_2^2 + \sigma_{23} = 0$$
$$\sigma_{13} + \sigma_{23} + \sigma_3^2 = 0 .$$

In case of ABO blood groups since there is dominance relationship between the alleles A, B and O, replacing the A_1, A_2 and A_3 alleles by A, B and O respectively one has the proportions of the four phenotypes as

$$O \; . \; . \; . \; . \; . \; . \quad r^2 + \sigma_r^2$$
$$A \; . \; . \; . \; . \; . \; . \quad p^2 + 2pr + \sigma_p^2 + 2\sigma_{pr}$$
$$B \; . \; . \; . \; . \; . \; . \quad q^2 + 2qr + \sigma_q^2 + 2\sigma_{qr}$$

and AB . . . . . . $2pq + 2\ \sigma_{pq}$.

The variances and the covariances, in this case, satisfy the relations :

$$2\,\sigma_{pq} = \sigma_r^2 - \sigma_p^2 - \sigma_q^2$$

$$2\,\sigma_{pr} = \sigma_q^2 - \sigma_p^2 - \sigma_r^2$$

$$2\,\sigma_{qr} = \sigma_p^2 - \sigma_q^2 - \sigma_r^2$$

Again remembering that p, q and r are the averages of the quantities lieing between 0 and 1, one has

$$\sigma_p^2 \leq p(1 - p),\ \sigma_q^2 \leq q(1 - q) \text{ and } \sigma_{pq} \leq p(1 - q) \text{ or } q(1 - p)$$

Now put

$$\begin{aligned} r_* &= \sqrt{r^2 + \sigma_r^2} \\ p_* &= \sqrt{(p + r)^2 + \sigma^2_{(p+r)}} - \sqrt{r^2 + \sigma_r^2} \\ &= \sqrt{(p + r)^2 + \sigma_q^2} - \sqrt{r^2 + \sigma_r^2} \\ \text{and}\quad q_* &= \sqrt{(q + r)^2 + \sigma^2_{(q+r)}} - \sqrt{r^2 + \sigma_r^2} \\ &= \sqrt{(q + r)^2 + \sigma_p^2} - \sqrt{r^2 + \sigma_r^2} \end{aligned} \tag{4.1.24}$$

One gets from (4.1.24),

$$r_*^2 = r^2 + \sigma_r^2$$

$$p_*^2 + 2p_* r_* = p^2 + 2pr + \sigma_p^2 + 2\sigma_{pr}$$

and $$q_*^2 + 2q_* r_* = q^2 + 2qr + \sigma_q^2 + 2\sigma_{qr}$$

Now if $p_* + q_* + r_* = 1$, then

$$\begin{aligned} 2p_* q_* &= 1 - (p_*^2 + q_*^2 + r_*^2 + 2p_* r_* + 2q_* r_*) \\ &= 1 - [(p_*^2 + 2p_* r_*) + (q_*^2 + 2q_* r_*) + (r^2 + \sigma_r^2)] \\ &= 2pq + 2\sigma_{pq} \quad [\text{since } \sigma^2_{(p+q+r)} = 0]\,. \end{aligned}$$

Conversely, if $2p_* q_* = 2pq + 2\sigma_{pq}$, then $p_* + q_* + r_* = 1$.

Thus we complete the proof of the theorem stated as follows:

Theorem 4.1 A mixture of random mating races with average A, B, and O gene frequencies p, q and r respectively cannot be distinguished from a homogeneous random mating population with gene frequencies p_*, q_* and r_* respectively, if and only if $p_* + q_* + r_* = 1$

where, $p_* = \sqrt{(p+r)^2 + \sigma_q^2} - \sqrt{r^2 + \sigma_r^2}$, $q_* = \sqrt{(q+r)^2 + \sigma_p^2} - \sqrt{r^2 + \sigma_r^2}$

and $r_* = \sqrt{r^2 + \sigma_r^2}$.

To make the picture a more concrete one let us consider the following example. Consider a population which is, in fact, a mixture of two mendelian isolates where the A, B, O gene frequencies are 0.0300, 0.1688, 0.8012 and 0.5700, 0.1488, 0.2812 respectively. Assuming that the relative sizes of these two isolates

are 0.5 and 0.5, one has the averages A, B, O gene frequencies for the whole population as

$$p = \frac{0.03 + 0.57}{2} = 0.3000$$

$$q = \frac{0.1688 + 0.1488}{2} = 0.1588$$

$$\text{and} \quad r = \frac{0.8012 + 0.2812}{2} = 0.5412 \ .$$

But, without knowing exactly this fact, if we proceed to obtain the gene frequencies by Wiener's formula, we obtain $p_* = 0.25$, $q_* = 0.15$ and $r_* = 0.60$ under the assumption of random mating. Thus a mixture of such two isolates will depict the same phenotypic frequencies as depicted by a homogeneous random mating population with, of course, different gene ratios.

4.1.4 Model II population and estimation of parameters :

The phenotype frequencies of the four blood groups O, A, B and AB for a Model II population are already shown in Table 4.2. In this case, thus, we have three linearly independent parameters, namely, p, q and F (r being related through $r = 1 - p - q$). There had been lot of controversies regarding the estimation of these parameters. In particular the question of reliability of the estimates (even the asymptotically most efficient maximum likelihood estimates) is still a debatable topic. Schull (1965) appears to be the first author to give the maximum likelihood estimates of p, q and F explicitly. Later on Schull and Ito (1969) have shown that the maximum likelihood estimates of p, q, r and F are obtainable by solving simultaneously

$$r^2 + r(1 - r)F = \overline{O},\ p^2 + 2pr + [p(1 - p) - 2pr]F = \overline{A},$$

$$q^2 + 2qr + [q(1 - q) - 2qr]F = \overline{B},\ 2pq(1 - F) = \overline{AB}$$

where $\overline{O}$, $\overline{A}$, $\overline{B}$ and $\overline{AB}$ are the observed relative frequencies of the phenotypes O, A, B and AB. Solution of the above set of simultaneous equation yields :

$$\hat{p} = \left\{ Z \pm [Z^2 - 8(\overline{AB} + 2\overline{B})(2\overline{A}.\ \overline{AB} + \overline{AB}^2)]^{\frac{1}{2}} \right\} / [4(\overline{AB} + 2\overline{B})]$$

$$\hat{q} = [\overline{AB}(1 - p)] / [2(\overline{A} - \hat{p} + \overline{AB})],\quad \hat{r} = 1 - \hat{p} - \hat{q}$$

$$\text{and}\quad \hat{F} = (\overline{O} - \hat{r}^2) / (\hat{r} - \hat{r}^2) \tag{4.1.25}$$

where $Z = 4\,\bar{A}\,.\,\overline{AB} + 4\,\bar{A}.\bar{B} + 4\,\bar{B}\,.\,\overline{AB} + 3\,\overline{AB}^{2}$.

Though these are the explicit expressions of the maximum likelihood estimates, in order to obtain the asymptotic variances of the estimates one has to follow the alternative standard technique wherein the estimates of p, q, r and F are derived by setting the partial derivatives with respect to p, q and F of the log likelihood function equal to zero, and solving simultaneously the resulting set of equations. Thus we have to solve the set:

$$\left.\begin{aligned}
U_p &= \frac{\overline{AB}}{p} - \frac{\bar{O}[2r + (1-2r)F]}{r^2 + r(1-r)F} + \frac{\bar{A}[2r + (1-2r)F]}{p^2+2pr+[p(1-p) - 2pr]F} \\
&\qquad - \frac{\bar{B}[2q(1-F)]}{q^2+2qr+[q(1-q) - 2qr]F} = 0\,, \\
U_q &= \frac{\overline{AB}}{q} - \frac{\bar{O}[2r + (1-2r)F]}{r^2 + r(1-r)F} - \frac{\bar{A}[2p(1-F)]}{p^2+2pr+[p(1-p) - 2pr]F} \\
&\qquad + \frac{\bar{B}[2r + (1-2r)F]}{q^2 + 2qr + [q(1-q) - 2qr]F} = 0, \\
U_F &= \frac{\bar{O}(1-r)}{r + (1-r)F} + \frac{\bar{A}(q-r)}{p + 2r + (q-r)F} + \frac{\bar{B}(q-r)}{q + 2r + (p-r)F} \\
&\qquad - \frac{\overline{AB}}{1-F} = 0
\end{aligned}\right\} \qquad (4.1.26)$$

Normally, one would resort to an iterative procedure to solve these equations using the information matrix with respect to p, q and F.

Yasuda (1966a, b) has shown that U_p, U_q and U_F when evaluated at F = 0 are related by

$$U_F = -\left(\frac{p}{2}\right) U_p - \left(\frac{q}{2}\right) U_q$$

which indirectly proves that the information matrix is singular and hence cannot be inverted when F = 0. We note that from (4.1.26), in general,

$$U_F + \left(\frac{p}{2}\right) U_p + \left(\frac{q}{2}\right) U_q \rightarrow 0 \quad \text{as} \quad F \rightarrow 0 \qquad (4.1.27)$$

and the efficient scores U_F, U_p, U_q are all continuous throughout the parametric range $(0 \leq p, q, F \leq 1)$. Schull and Ito thought that U_F is not continuous at F = 0. (One may note that the observation regarding the continuity of U_F at F = 0 is also noted by Yee and Morton (1970) in a letter in <u>Amer. J. Hum. Genet.</u>).

From the above points it is clear that (1) At F = 0, since the information matrix is singular one cannot adopt the iterative procedure and hence the asymptotic expressions of the estimates cannot be obtained, (2) When F is very small (at the levels of F which can be observed in human populations), even in very large samples the ABO and similar systems provide negligible information about F, (3) Working out the explicit expressions of the elements of the matrix, one can notice that determinant of the information matrix converges to zero as F → 0 (which indirectly follows from property (4.1.27)). This shows that asymptotic generalized variannce of the maximum likelihood estimates is large enough as to question the reliability of the estimates.

Yee and Morton (1970), through an empirical study, has also shown that the standard error of the estimate of F is enormous but for the extreme values of F which are absurd so far as human populations are concerned. Their empirical works suggest that it is very difficult to distinguish between Model I and Model II population for realistic values of F. For a ready verification of these two facts we reproduce their results in Table 4.3

TABLE 4.3

Maximum likelihood estimation of F for ABO system
($p = 0.22$, $q = 0.16$, $r = 0.62$)

	F						
	-0.5	-0.05	-0.005	0	0.005	0.05	0.5
χ^2	637.00	0.07	0	0	0	0.07	942.00
σ_F	0.012	0.097	0.940	∞	0.930	0.090	0.006

Source : Yee and Morton (1970).

Moreover, for the general case of multiple alleles, it has been argued (see Li and Horvitz, 1953) that since the method of maximum likelihood does not yield the usual gene frequency estimates (this being true in the present case also, which can be verified through (4.1.25)) it may be best to accept the conventional values and estimate F under this set of conditions. As an example, let us consider a sample of 115 individuals of whom 49 are of type A, 35 of type B, 24 of type O, and 7 of type AB. Equations in (4.1.25) lead to an estimate of F of 0.7781, whereas the other leads to an estimate of approximately 0.00005. A direct estimate of F within this population gives rise to a value of about 0.006 (Schull, Yanase and Nemoto, 1962). Thus, if one accepts this latter estimate as the "true" value, then both methods would appear to be off by a factor of 120 or so, but in different directions.

All these only suggest that for human populations (wherein, generally, F is in the vicinity of zero) some other estimation procedures are to be explored, especially, for estimating the so called inbreeding coefficient, F, from ABO phenotypic bioassay though it is clear that other data structures may require or profit from other methods of analysis.

Detection of F from ABO blood group data :

We have already stated that Yee and Morton's empirical works suggested that even for a large sample the presence of F cannot be detected but for unrealistically large value (like F = ± 0.5). Herein, we study the problem a bit more analytically and tabulate the minimum sample size required for different levels of F-values. The problem is tackled by a consideration of the power of a χ^2 - test.

If we denote the O, A, B and AB phenotype frequencies under Model II by π_1, π_2, π_3 and π_4 respectively, then we have

$$\pi_i = \pi_i^* + \frac{C_i}{\sqrt{n}}$$

where π_i^*'s are the corresponding proportions under the Model I population structure (null hypothesis F = 0) and

$$C_1 = Fr(1 - r) \sqrt{n}$$

$$C_2 = Fp(q - r) \sqrt{n}$$

$$C_3 = Fq(p - r) \sqrt{n}$$

and $$C_4 = - 2Fpq \sqrt{n}$$

n, being the sample size.

To detect any deviation ($F \neq 0$) from Hardy-Weinberg proportions (Model I) one employs a non-central χ^2 - test (see Chapman, 1968) whose non-centrality parameter, λ, is given by

$$\lambda = \Sigma \left[c_i^2 / \pi_i^* \right]$$

$$= n F^2 \left[2pq + (1 - r)^2 + \frac{p(q - r)^2}{p + 2r} + \frac{q(p - r)^2}{q + 2r} \right].$$

From this expression the minimum sample size, n, can also be determined to detect a specified level of inbreeding coefficient by a α-level test procedure with power β. Notice that λ, the non-centrality parameter is completely specified once we know the level and power of the test procedure (Owen, 1962; Johnson and Pearson, 1969).

Table 4.4 presents the minimum sample size, n, required to detect a specified level of inbreeding coefficient for different p, q combinations. It can be noticed that for levels of F, which are observed in natural populations (i.e., less than 5 per cent) the required sample size is hopelessly large which indicates the absurdity of the detection of F through this analytical procedure. It is worth mentioning that such a treatment is also given by Ward and Sing (1970) wherein they considered the simple cases where the genes at the locus of interest have no dominance relationship between them. In such a case, of course, the non-centrality parameter λ, and thereby the sample size required, does not depend upon the relative frequencies of the genes concerned. The similarity of such a case with genetic correlations (see Chapter I) is worth observing which indirectly suggests a possibility of detection of F through the genetic correlations.

TABLE 4.4

Sample sizes to obtain a specified power β, using an α size test procedure for vaiious values of p and q.

($\alpha = 0.01$)

	$p = 0.15$, $q = 0.15$		
F	$\beta = 0.2$	$\beta = 0.5$	$\beta = 0.9$
0.0001	1555166667	3430666668	7688000003
0.0005	62206666	137226666	307520000
0.0010	15551666	34306666	76880000
0.0050	622066	1372266	3075200
0.0100	155516	343066	768800
0.0500	6220	13722	30752
0.1000	1555	3430	7688
0.2500	248	548	1230
0.5000	62	137	307
1.0000	15	34	76

TABLE 4.4 (continued)

F	$\beta = 0.2$	$\beta = 0.5$	$\beta = 0.9$
		$p = 0.15,\ q = 0.15$	
0.0001	1053500000	2324000000	5208000000
0.0005	42140000	92960000	208320000
0.0010	10535000	23240000	52080000
0.0050	421400	929600	2083200
0.0100	105350	232400	520800
0.0500	4214	9296	20832
0.1000	1053	2324	5208
0.2500	168	371	833
0.5000	42	92	208
1.0000	10	23	52
		$p = 0.20,\ q = 0.25$	
0.0001	887823529	1958521008	4388974790
0.0005	35512941	78340840	175558991
0.0010	8878235	19585210	43889747
0.0050	355129	783408	1755589
0.0100	88782	195852	438897
0.0500	3551	7834	17555
0.1000	887	1958	4388
0.2500	142	313	702
0.5000	35	78	175
1.0000	8	19	43

TABLE 4.4 (Continued)

	p = 0.25, q = 0.25		
F	β = 0.2	β = 0.5	β = 0.9
0.0001	752500000	1660000000	3720000000
0.0005	30100000	66400000	148800000
0.0010	7525000	16600000	37200000
0.0050	301000	664000	1488000
0.0100	75250	166000	372000
0.0500	3010	6640	14880
0.1000	752	1660	3720
0.2500	120	265	595
0.5000	30	66	148
1.0000	7	16	37

p = 0.15, q = 0.15

0.0001	640666666	1984000000	5430166669
0.0005	25626666	79360000	217206666
0.0010	6406666	19840000	54301666
0.0050	256266	793600	2172066
0.0100	64066	198400	543016
0.0500	2562	7936	21720
0.1000	640	1984	5430
0.2500	102	317	868
0.5000	25	79	217
1.0000	6	19	54

TABLE 4.4 (Continued)

		p = 0.20, q = 0.20	
F	β = 0.2	β = 0.5	β = 0.9
0.0001	434000000	1344000000	3678500000
0.0005	17360000	53760000	147140000
0.0010	4340000	13440000	36785000
0.0050	173600	537600	1471400
0.0100	43400	134400	367850
0.0500	1736	5376	14714
0.1000	434	1344	3678
0.2500	69	215	588
0.5000	17	53	147
1.0000	4	13	36
		p = 0.20, q = 0.25	
0.0001	365747899	1132638655	3100008403
0.0005	14629915	45305546	124000336
0.0010	3657478	11326386	31000084
0.0050	146299	453055	1240003
0.0100	36574	113263	310000
0.0500	1462	4530	12400
0.1000	365	1132	3100
0.2500	58	181	496
0.5000	14	45	124
1.0000	3	11	31

TABLE 4.4 (Continued)

	p = 0.25, q = 0.25		
F	$\beta = 0.2$	$\beta = 0.5$	$\beta = 0.9$
0.0001	310000000	960000000	2627500000
0.0005	12400000	38400000	105100000
0.0010	3100000	9600000	26275000
0.0050	124000	384000	1051000
0.0100	31000	96000	262750
0.0500	1240	3840	10510
0.1000	310	960	2627
0.2500	49	153	420
0.5000	12	38	105
1.0000	3	9	26

4.2 MNSs BLOOD GROUP SYSTEM

4.2.0 Genetics of MNSs blood groups :

The MNSs system can best be described in terms of two loci on a single chromosome and very closely linked. One locus is occupied by either M or N, the other by S or s. Each of the genes M, N, S and s gives rise to a corresponding antigen, recognisable by agglutination of the red cells with the appropriate antibody.

There are thus three MN phenotypes M, MN and N, corresponding respectively to the genotypes MM, MN and NN. It is in terms of these that most anthropological works have been done. The discovery of S doubles the number of phenotypes and increases the number of genotypes to ten, as shown in Table 4.2.1. Anti-s, being the rarest antibody of this lot, is generally not available for anthropological works.

TABLE 4.2.1

Genotypes and Phenotypes in MNSs system

Phenotypes (Tested with anti-M, -N, -S)	Phenotypes (Tested with anti-M, -N, -S, -s)	Genotypes
M	Ms	MsMs
MS	MS	MSMS
	MSs	MSMs
MN	MNs	MsMs
MNS	MNS	MSNS
	MNSs	MSNs
		MsNS
N	Ns	NsNs
NS	NS	NSNS
	NSs	NSNs

4.2.1 Estimation of chromosome frequencies under Model I population Structure: phenotypes detected by three sera :

As the Table 4.2.1 indicates, if anti-s serum is not used, one can only identify six distinct phenotypes and one may note that the systems involves four different chromosome structure, namely, Ms, MS, Ns and NS. We denote the relative frequencies of these chromosomes by m_s, m_S, n_s and n_S, and the observed proportions of individuals in each of the six phenotypes by $\overline{M}$, $\overline{MS}$, $\overline{MN}$, $\overline{MNS}$, $\overline{N}$, and $\overline{NS}$. The expected values of these proportions, under Model I population structure, are

$$
\begin{aligned}
E(\overline{M}) &= m_s^2 & E(\overline{MNS}) &= 2(m_s n_S + m_S n_s + m_S n_S) \\
E(\overline{MS}) &= m_S^2 + 2\, m_S m_s & E(\overline{N}) &= n_s^2 \qquad\qquad (4.2.1 \\
E(\overline{MN}) &= 2\, m_s n_s & E(\overline{NS}) &= n_S^2 + 2\, n_S n_s
\end{aligned}
$$

It is easy to note that $m = m_S + m_s$ represents the frequency of the M-gene if the other locus (occupied by S or s) is completely ignored. Similarly, $n = n_S + n_s$ represents the N gene frequency, $s = m_s + n_s$, the s-gene frequency and $S = m_S + n_S$ the S-gene frequency. Note that $m_S + m_s + n_S + n_s = 1$.

m is estimated by the usual 'gene count' estimate

$$m' = \overline{M} + \overline{MS} + \frac{\overline{MN}}{2} + \frac{\overline{MNS}}{2} \qquad (4.2.2)$$

and hence the estimate of n is $n' = 1 - m'$. These are also the maximum lilelihood estimates (see DeGroot, 1956) having variances as

$$V(m') = V(n') = \frac{mn}{2G} \qquad (4.2.3)$$

where G is the total number of individuals sampled. Mourant (1954, pp. 220) proceeded to estimate s as

$$s' = \sqrt{\overline{M} + \overline{MN} + \overline{N}} \qquad (4.2.4)$$

the explanation of which is evident from the expressions in (4.2.1) and thus he obtained

$$m'_s = m' \sqrt{\overline{M} / (\overline{MS} + \overline{M})} \qquad (4.2.5)$$

$$n'_s = n' \sqrt{\overline{N} / (\overline{NS} + \overline{N})}.$$

It is to be noted here that $m'_s + n'_s$ need not add to s'. The next step is to adjust them so that they do so. The adjustment was done as

$$m''_s = \frac{m'_s \; s'}{m'_s + n'_s} , \quad n''_s = \frac{n'_s \; s'}{m'_s + n'_s} .$$

In terms of the observed phenotypic proportions

$$m''_s = \frac{(2M+H)\ \sqrt{\overline{M}/M}\ \ \sqrt{R}}{(2M+H)\sqrt{\overline{M}/M} + (H+2N)\ \sqrt{\overline{N}/N}} \qquad (4.2.6)$$

where, $M = \overline{MS} + \overline{M}$, $H = \overline{MNS} + \overline{MN}$, $N = \overline{NS} + \overline{N}$

and $R = \overline{M} + \overline{MN} + \overline{N}$.

Its variance is unknown and difficult to obtain.

A much simpler method of estimating m_s and n_s has been suggested by Wiener (1954). The estimate for m_2 , for instance, is

$$m'_s = \tfrac{1}{2} \left[\sqrt{\overline{M} + \overline{MN} + \overline{N}} + \sqrt{\overline{M}} - \sqrt{\overline{N}} \right] \qquad (4.2.7)$$

Boyd (1956) obtained its variance as

$$V(m'_s) = \frac{1 - 4\ m_s^2 + 4\ m_s/s}{16\ G} \qquad (4.2.8)$$

Boyd (1955) discussed the simplifications of mathematics resulting from a reparametrization given as $g = m_s/m$ and $d = n_s/n$. We can interpret g as the conditional probability that a chromosome is m_s given that a M gene is there and d can be interpreted analogously. Note that with this one has to estimate only three linearly independent parameters m, g and d, all lieing between zero and one. Even with this reparametrization the maximum likelihood method is a long and time-consuming one. But since maximum likelihood method provides fully efficient estimates, it is good to know the asymptotic variances of such estimates for one can use them to evaluate the efficiencies of the other simpler methods. DeGroot (1956) made a study of the covariance structure of the m.l. estimates. He obtained the variances and covariances of g' and d' as

$$\left.\begin{aligned} V(g') &= \frac{g(1 - g^2)(2dn + gm - ngd^2 - gd^2)}{4Gm\left[(gm + dn)^2 - g^2d^2\right]} \\ V(d') &= \frac{d(1 - d^2)(2gm + dn - mdg^2 - dg^2)}{4Gn\left[(gm + dn)^2 - g^2d^2\right]} \\ CV(g',d') &= \frac{-gd(1 - g^2)(1 - d^2)}{4G\left[(gm + dn)^2 - g^2d^2\right]} \end{aligned}\right\} \qquad (4.2.9)$$

Finally, for the variances and covariances involving $m' + n'$, he had

$$\left.\begin{aligned} V(n') &= V(m') \\ CV(m, n') &= - V(m') \\ CV(n', d') &= CV(n', g') = 0 \\ CV(m', g') &= CV(m', d') = 0. \end{aligned}\right\} \qquad (4.2.10)$$

In terms of (4.2.9) and (4.2.10), DeGroot wrote down the expressions of the variances and the covariances of the estimates of the chromosome frequencies.

DeGroot and Li (1960) presented yet another method which is computationally much simpler and yet almost as efficient as the maximum likelihood method. Like Mourant (1954) they also estimated m and s by means of the expressions (4.2.2) and (4.2.4) and then m_s and n_s were estimated by

$$m_s' = \left(\frac{2\,\overline{M} + \overline{MN}}{R}\right) s' \quad \text{and} \quad n_s' = \left(\frac{2\,\overline{N} + \overline{MN}}{R}\right) s' \qquad (4.2.11)$$

Variances of these estimates were obtained as

$$\left.\begin{aligned}
V(s') &= \frac{1 - s^2}{4\,G} \\
V(m_s') &= \frac{m_s\,n_s}{2G\,s^2} + \frac{m_s^2}{s^2} + \frac{m_s^2}{s^2} \cdot \frac{1 - s^2}{4G} \\
V(n_s') &= \frac{m_s\,n_s}{2G\,s^2} + \frac{n_s^2}{s^2} \cdot \frac{1 - s^2}{4G} \\
V(m_S') &= V(m_s') + \frac{n(m_S - m_s)}{2G} \\
V(n'_S) &= V(n_s') + \frac{m(n_S - n_s)}{2G}
\end{aligned}\right\} \qquad (4.2.12)$$

From (4.2.8) and (4.2.12) it is evident that

$$\text{Wiener's } V(m_s') = \text{DeGroot's } V(m_s') + \frac{(m_s - n_s)^2}{16\,G\,s^2}$$

and thus unless $m_s = n_s$, DeGroot and Li's estimates are more efficient than those of Wiener's. However, such a general statement should be made only after comparing the generalized variance of the estimates analytic solution of which is yet to be in literature. Analytic comparison of these methods with the maximum likelihood one also does not appear to be published till now. DeGroot and Li, however, gave an illustration to show that the standard errors of the estimates obtained by their method, though larger, but very close to those of the maximum likelihood estimates.

4.2.2 Estimation of chromosome frequencies under Moder I population structure : phenotypes detected by 4 sera :

Use of anti-s serum enables us to identify as many as 9 phenotypes. Only the genotypes MSNs and MsNS are not distinguishable (see Table 4.2.1). Because of the heavy amount of labour involved for computing the maximum likelihood estimates in certain cases, particularly when it is sufficient to work with an estimate considerably close to the m.l.e's, it is necessary to get a quicker estimate which may not be that efficient as the one already told about. Smith (1957,

1967) discussed the gene-count method in general for multi-allelic systems. Herein we develop a counting method to obtain the estimate of the chromosome frequencies. Though, strictly speaking, the estimates thus obtained are inefficient, but advantage of this method lies in its computational simplicity.

In a sample of G unrelated individuals from a Model I population, let the observed and expected frequencies of the 9 phenotypes are given as in Table 4.2.2.

Now, the estimate of m is easily obtained by counting as

$$m' = \frac{E_1}{2G} = \frac{2n_1 + 2n_2 + 2n_3 + n_4 + n_5 + n_6}{2G} \qquad (4.2.13)$$

where E_1 = Number of M-gene in the sample.

and so $n' = 1 - m'$.

The variance of m' and n' are same as in the expression (4.2.3).

Estimation of g and d : Let E_3 and E_4 denote the number of Ms and Ns chromosome in the sample respectively. Now from Ms, MSs and Ns individuals we get a contribution of $n_2 + 2n_3 + n_6$ to E_3 and from Ns, NSs and MNs individuals the contribution being

$n_6 + n_8 + 2n_9$ to E_4.

From individuals of type MNSs we have the contributions to E_3 and E_4 in the ratio

$$\frac{g(1-d)}{g(1-d) + d(1-g)} : \frac{d(1-g)}{g(1-d) + d(1-g)}$$

Thus, from the whole sample

$$\left.\begin{aligned} E_3 &= n_2 + 2n_3 + n_6 + n_5 \cdot \frac{g(1-d)}{g(1-d) + d(1-g)} \\ E_4 &= n_6 + n_8 + 2n_9 + n_5 \cdot \frac{d(1-g)}{g(1-d) + d(1-g)} \end{aligned}\right\} \qquad (4.2.14$$

These expressions being dependent on g and d cannot be computed unless we have some provisional estimates of g and d. Fairly good provisional estimates of g and d are given by

$$g_{(1)} = \sqrt{\frac{n_3}{n_1 + n_2 + n_3}} \quad \text{and} \quad d_{(1)} = \sqrt{\frac{n_9}{n_7 + n_8 + n_9}}$$

Using these we get the chromosome count of Ms and Ns by formulae (4.2.14).

The improved estimates of g and d are given by

TABLE 4.2.2

Observed and Expected frequencies for all phenotypes for MNSs system under random mating

Phenotype	Observed Frequencies	Expected proportions	
MS	n_1	m_S^2	$m^2(1-g)^2$
MSs	n_2	$2m_Sm_s$	$2m^2g(1-g)$
Ms	n_3	m_s^2	$m^2 g^2$
MNS	n_4	$2m_Sn_S$	$2mn(1-g)(1-d)$
MNSs	n_5	$2(m_Sn_s+m_sn_S)$	$2mn(g+d-2gd)$
MNs	n_6	$2m_sn_s$	$2mngd$
NS	n_7	n_S^2	$n^2(1-d)^2$
NSs	n_8	$2n_Sn_s$	$2n^2d(1-d)$
Ns	n_9	n_s^2	$n^2 d^2$
Total	G	1	1

$$g_{(2)} = \frac{E_3(1)}{E_1} \quad \text{and} \quad d_{(2)} = \frac{E_4(1)}{E_2}$$

where $E_3(1)$ = Number of Ms chromosome in the sample (with provisional estimates $g_{(1)}$ and $d_{(1)}$).

$E_4(1)$ = Number of MS chromosome in the sample (with provisional estimates $g_{(1)}$ and $d_{(1)}$).

$= E_1 - E_3(1)$.

$E_5(1)$ = Number of Ns chromosome in the sample (with provisional estimates $g_{(1)}$ and $d_{(1)}$).

and $E_6(1)$ = Number of NS chromosome in the sample (with provisional estimates $g_{(1)}$ and $d_{(1)}$).

$= E_2 - E_5(1)$.

$E_1 + E_2$ being equal to 2G.

This process can be repeated unless g and d values are stable. But the number of steps can be reduced by using the ordinary maximum likelihood adjustments. This procedure involves the compuation of hidden vàriance matrix and the information matrix, detailed indication of which is given in Ceppellini et al. (1955) and Smith (1957).

Numerical Example

For illustration let us consider a sample which was tested with four sera by Race and Sanger (1951). The phenotypic frequencies being $n_1 = 18$, $n_2 = 45$, $n_3 = 19$, $n_4 = 5$, $n_5 = 45$, $n_6 = 51$, $n_7 = 2$, $n_8 = 11$ and $n_9 = 33$.

$$\text{So} \quad E_1 = 265, \quad E_2 = 193$$

$$2G = E_1 + E_2 = 458$$

Consequently,

$$m' = \frac{E_1}{2G} = 0.578603$$

$$\text{and} \quad n' = 1 - m' = 0.421397.$$

The estimate of their error - variances are given by

$$\frac{m'\,n'}{2G} = 0.000532.$$

The provisional estimates of g and d are

$$g_{(1)} = \sqrt{\frac{19}{18 + 45 + 19}} = 0.47135$$

$$\text{and} \quad d_{(1)} = \sqrt{\frac{33}{2 + 11 + 33}} = 0.84699$$

and hence

$$E_3(1) = 140.24267$$

$$E_4(1) = 124.75733$$

$$E_5(1) = 166.75733$$

$$E_6(1) = 26.24267 \text{ .}$$

The improved estimates of g and d now are

$$g_{(2)} = \frac{140.24267}{265} = 0.52922$$

and

$$d_{(2)} = \frac{166.75733}{193} = 0.86403 \text{ .}$$

With these estimates $g_{(2)}$ and $d_{(2)}$ the total hidden covariance matrix is found to be

$$H = \begin{bmatrix} 5.74726 & -5.74726 & -5.74726 & 5.74726 \\ & 5.74726 & 5.74726 & -5.74726 \\ & & 5.74726 & -5.74726 \\ & & & 5.74726 \end{bmatrix}$$

from which the information matrix is obtained as :

$$J = \begin{bmatrix} 482.075 & 23.068 & 12.569 & -79.870 \\ \cdot & 534.615 & -14.129 & 89.784 \\ & & 214.975 & 48.920 \\ & & & 1136.775 \end{bmatrix}$$

In both of these matrices the columns as well as rows correspond to E_3, E_4, E_5 and E_6. From J we get the covariance matrix corresponding to the estimates of g and d as:

$$V^* = \begin{bmatrix} V_{33} & V_{35} \\ V_{53} & V_{55} \end{bmatrix} = \begin{bmatrix} 0.001064 & -0.000167 \\ -0.000167 & 0.000824 \end{bmatrix}$$

In addition we find the scores

$$U_3^* = \frac{E_3}{g} - \frac{E_4}{1 - g} = 2.0898$$

$$U_5^* = \frac{E_5}{d} - \frac{E_6}{1 - d} = -4.4410 .$$

The further improved values of g and d are then given by

$$g_{(3)} = g_{(2)} + V_{33} U_3^* + V_{35} U_5^* = 0.53219$$

$$d_{(3)} = d_{(2)} + V_{53} U_3^* + V_{55} U_5^* = 0.86002$$

The process is now repeated only changing the values of g, d and the E-values. The other things, in practice, can be kept the same. The eventual estimates are

$$g' = 0.53221$$

and $$d' = 0.85992 \ .$$

The estimates g', d' have error-variances equal to V_{33} and V_{55} (at final estimates) respectively. In particular,

$$V(g') = 0.001069$$

and $$V(d') = 0.000867.$$

Now we obtain the estimates of the chromosome frequencies and their error-variances as

$$m'_s = m' \ g' = 0.307938$$

$$m'_S = m' - m' \ g' = 0.270665$$

$$n'_s = n' \ d' = 0.362368$$

$$n'_S = n' - n' \ d' = 0.059029$$

and $$V(m'_s) = g'^2 \ V(m') + m'^2 \ V(g') = 0.000509$$

$$v(m'_S) = (1 - g')^2 \ V(m') + m'^2 \ V(g') = 0.000474$$

$$V(n_s') = d'^2\, V(n') + n'^2\, V(d') = 0.000547$$

$$V(n_S') = (1 - d')^2\, V(n') + n'^2\, V(d') = 0.000164 .$$

At last we compare these estimates with those obtained by Rao (1969) with the same data. Table 4.2.3 gives the estimates obtained by the two methods. The comparison shows that the estimates of the chromosome frequencies agree more or less with the maximum likelihood estimates but the error-variances are slightly but consistently larger, except for n_S - chromosome.

TABLE 4.2.3

Comparison of estimates and their variances with max. likelihood method

Chromosome frequencies	Gene Count method		Max. likelihood method	
m_s	0.307938	0.000509	0.307938	0.000371
m_S	0.270665	0.000474	0.270665	0.000337
n_s	0.362368	0.000547	0.362367	0.000167
n_S	0.059029	0.000164	0.059030	0.000216

4.2.3 Estimation.of chromosome frequencies under more general set up :

Recently there have been attempts to obtain the chromosome frequencies under Model II population structure (Rao, 1969). But since the problem of reliability of the estimate of F is very much there we are not going to discuss the method here. The validity of the maximum likelihood estimates for populations with values of F in the viccinity of zero is still to be explored.

However, the use of the restricted random mating model can be made with profit since the reliability of such estimates are not subjected to any question. The actual estimation procedure, being a routine application of the methodology developed in Chapter II, is not presented here again. As a last word, it can only be mentioned that these estimates are also fully efficient.

REFERENCES

ADAM, A.; SHEBA, C.: SANGER, R.; RACE, R. R.,; TIPPETT, P.; HAMPER, J.; GAVIN, J. and FINNEY, D. J. (1963). Data for X-mapping calculations, Israeli families tested for Xg, g-6-pd and for colour vision. Ann. Hum. Genet., London, 26, 187-194.

ADAM, A.; TIPPETT, P.; GAVIN, J.; NOADES, J.; SANGER, R. and RACE, R. R. (1967). The linkage relation of Xg to g-6-pd in Israelis: the evidence of a second series of families. Ann. Hum. Genet., Lond., 30, 211-218.

ALLEN, G. (1960). A differential method of estimation of type frequencies in triplets and quadruplets. Amer. J. Hum. Genet., 12, 210-224.

BERSTEIN, F. (1925). Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen. Zeitschr. Abstags. u.Vererbgsl., 37, 237-270.

BERNSTEIN, F. (1930). Fortgesetzte Untersuchungen aus der Teorie der Blutgruppen. Z. indukt. Abstamm. Vererb. Lehr., 56, 223-273.

BOORMAN, K. E. (1950). An analysis of the blood types and clinical condition of 2000 consecutive mothers and their infants. Ann. Eugen. Lond., 15, 120-134.

BOYD, W. C. (1954). Gene frequencies in Anthropology : simple methods. Amer. J. Phys. Anthrop., 12, 241-251.

BOYD, W. C. (1955). Letter to the editor. Amer. J. Hum. Genet., 7, 444-445.

BOYD, W. C. (1956). Variances of gene frequency estimates. Amer. J. Hum. Genet., 8, 24-38.

BOYE, C. L. (1941). An allelic series in Coleus. J. Genet., 42, 191.

BULMER, M. G. (1960). The familial incidence of twinning. Ann. Hum. Genet., Lond., 24, 1-3.

BULMER, M. G. (1958). The numbers of human multiple births. Ann. Hum. Genet. Lond., 22, 158-164

CEPELLINI, R.; SINISCALCO, M. and SMITH, C. A. B. (1955). The estimation of gene frequencies in a random mating population. Ann. Hum. Genet., Lond., 20, 97-115.

CHAKRABORTY, R. (1970a). Parent offspring correlation in an equilibrium population. To appear in Amer. J. Hum. Genet.

CHAKRABORTY, R. (1970b). Further results on parent offspring correlation in an equilibrium population. To appear in Amer. J. Hum. Genet.

CHAKRABORTY, R. (1970c). Genetic correlations in an equilibrium population. To appear in Sankhyā, B.

CHAKRABORTY, R. (1970d). A note on parent offspring correlation and inbreeding. To appear in Acta Genet. et Gamell.

CHAKRABORTY, R. (1970e). Gene frequency estimates in ABO system and their efficiencies. To appear in Sankhyā, B.

CHAKRABORTY, R. (1970f). A note on gene count method of estimation for MNSs blood group system. Submitted to Bull. Cal. Stat. Association.

CHAPMAN, D. C. (1968). Counted data. In: International Encyclopedia of Social Sciences. MacMillan and Free Press, 3, 417-427.

COTTERMAN, C. W. (1953). Regular two-allele and three-allele phenotype systems. Part I. Amer. J. Hum. Genet., 5, 193.

DAHLBERG, G. (1926). Twin births and twins from a hereditary point of view. Stockholm. A. B. Tidens Tryckeri.

DAS, S. R. (1953). A mathematical analysis of the phenomena of human twins and higher plural births. Part I : Twins. Metron, 17, 1-24.

DAS, S. R. (1955). A mathematical analysis of the phenomena of human twins and higher plural births. Part II : Triplets and the application of the analysis in the interpretation of the twin and the triplet data. Metron, 17, 4-27.

DAS, S. R. (1956). A mathematical analysis of the phenomena of human twins and higher plural births. Part III. Metron, 18, 1-46.

DEGROOT, M. H. (1956). Efficiency of gene frequency estimates for the ABO system. Amer. J. Hum. Genet., 8, 39-43.

DEGROOT, M. H. (1956). The covariance structure of the maximum likelihood gene frequency estimates for the MNS system. Amer. J. Hum. Genet., 8, 229-235.

DEGROOT, M. H. and LI, C. C. (1960). Simplified method of estimating the MNS gene frequencies. Ann. Hum. Genet., London, 24, 109-115.

DOBSON, A. M. and IKIN, E. W. (1946). The ABO blood groups in the United Kingdom; frequencies based on a very large population. J. Path. Bact., 48, 221.

ENDERS, T. and STERN, C. (1948). Frequencies of twins relative to age of mothers in American populations. Genetics, 33, 263-272.

FELLER, W. (1962). An introduction to Probability theory and its applications. Vol. I. Wiley Eastern Pvt. Ltd.. New Delhi.

FELLER, W. (1969). An introduction to Probability theory and its applications. Vol. II. Wiley Eastern Pvt. Ltd., New Delhi.

FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc., Edinborough, 52, 399-433.

FISHER, R. A. and TAYLOR, G. L. (1940). Scandinavian Influence on Scottish ethnology. *Nature, 145,* 590.

FISHER, R. A. (1950). *Statistical methods for Research Workers.* 11th ed., Oliver and Boyd, Oxford.

GREULICH, W. W. (1935). Heredity in human twinning. *Amer. J. Phys. Anthrop., 19,* 391-481.

HOGBEN, L. (1933). A matrix notation for Mendelian populations. *Proc. Roy. Soc., Edinborough,* 53, 7-25.

HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika, 28,* 321-377.

JAIN, S. K. and WORKMAN, P. L. (1967). Generalized F statistics and the theory of inbreeding and selection. *Nature, 214,* 674-678.

JENKINS, R. L. (1927). The interrelations of the frequency of plural births. *J. Hered., 8,* 387 and 504.

JOHNSON, N. L. and PEARSON, E. S. (1969). Tables of percentage points of non-central χ^2. *Biometrika, 56,* 255-272.

KARLIN, S. (1968). Equilibrium behavior of population genetic models with non-random mating. Part II : Pedigrees, homozygosity and stochastic models. *J. Appl. Prob., 5,* 487-566.

KENDALL, D. G. (1948). On the generalized 'birth and death' process. *Ann. Math. Statist., 19,* 1-15.

KEMPTHORNE, O. (1957). *An introduction to genetic statistics.* John Wiley and Sons, New York.

KOMAI, T. (1953). Composition of wild populations in the Lycaenid butterfly Neozyphyrus taxila. Amer. Nat., 87, 87.

KUDO, A. (1962). A method for calculating the inbreeding coefficient. Amer. J. Hum. Genet., 14, 426-432.

LEVINE, P. (1954). Gene frequencies in nonexperimental populations. In : Statistics and Mathematics in Biology, ed., Kempthorne, O. et al., Iowa College Press, Ames, Iowa, pp. 449-465.

LI, C. C. and HORVITZ, D. G. (1953). Some methods of estimating the inbreeding coefficient. Amer. J. Hum. Genet., 5, 107-117.

LI, C. C. (1954). The correlation between parents and offspring in a random mating population. Amer. J. Hum. Genet., 6, 383-386.

LI, C. C. and SACKS, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. Biometrics, 10, 347-360.

LI, C. C. (1955). Population Genetics, Univ. of Chicago Press, Chicago.

LI, C. C. (1956). The components of sampling variance of the ABO gene frequency estimates. Amer. J. Human. Genet., 8, 133-137.

LI, C. C. (1968). Fisher, Wright and Path coefficients. Biometrics, 24, 471-483.

LI, C. C. (1969). Population subdivision with respect to multiple alleles. Ann. Hum. Genet., Lond., 33, 23-29.

MALÉCOT, G. (1948). Les Mathematique de l'heredite. Masson. Paris.

MATHER, K. (1951). The measurement of Linkage in heredity. Methuen and Co. Ltd., London.

MORTON, N. E. and YASUDA, N. (1962). The genetical structure of populations. In : Les deplacements humains, Entretien de Monaco Sciences humains. Sutter, J., ed., pp. 186-203.

MOURANT, A. E. (1954). The distribution of the human blood groups. Blackwell Scientific Publications, Oxford.

MOURANT, A. E.; KOPÉC, A. C.; SOBCZAK, K. D. (1958). The ABO blood groups: Comprehensive tables and maps of world distribution. Blackwell Scientific Publications, Oxford.

NEEL, J. V. and SCHULL, W. J. (1954). Human heredity. Univ. of Chicago Press, Chicago.

NEWMAN, H. H. (1940). Multiple human births : twins, triplets, quadruplets and quintuplets. Doubleday, Doran and Co., New York.

OWEN, D. B. (1962). Handbook of Statistical tables. Addision-Wesley, Reading, Mass.

RACE, R. R.; IKIN, E. W.; TAYLOR, G. L. and PRIOR, A. M. (1942). A second series of families examined in England for A_1A_2BO and MN blood group factors. Ann. Eugen., London, 11, 385-394.

RACE, R. R.; SANGER, R.; LAWLER, S. D. and BERTINSHAW, D. (1949). The inheritance of the MNS blood groups: A second series of families. Heredity, 3, 205-213.

RACE, R. R. and SANGER, R. (1950). Blood groups in man. Blackwell Scientific Publications, Oxford.

RACE, R. R. and SANGER, R. (1951). The MNSs blood group system. Amer. J. Hum. Genet., 3, 322-343.

RAO, D. C. (1970). Statistical Methods in Blood groups. In : Proceedings of the second MASTECH Conference, Madras, India.

RAO, C. R. (1952). Advanced Statistical Methods in Biometric Research. John Wiley and Sons, New York.

RAO, C. R. (1965). Linear Statistical Inference and its Applications. John Wiley and Sons, New York.

ROSS, W. F. (1952). Twin frequency in-the Africans, Brit. Med. Journ., 2, 1336-1337.

SANGER, R.; RACE, R. R.; WALSH, R. J. and MONTGOMERY, C. (1948). An antibody which subdivides the human MN blood groups. Heredity, 2, 131-139.

SARKAR, S. S. (1944). The frequency of multiple births in India and Ceylon. Trans. Bose Res. Inst., 15.

SARMA, Y. R. and CHAKRABORTY, R. (1970). Some statistical models for human multiple births. Proc. second MASTECH Conference, Madras, India.

SCHULL, W. J.; Yanase, T. and NEMOTO, H. (1962). Kuroshima: The impact of religion on an Island's genetic heritage. Hum. Biol., 34, 271-298.

SCHULL, W. J.; ITO, P. K. and SONI, A. (1963). A note on Inbreeding and the Island of Susak. In : Jugoslavenska Akademija Znanosti i Umjetnosti.

SCHULL, W. J. (1965). Estimation of genetic parameters in population studies. In : Genetics and the epidemiology of chronic diseases. Neel, J. V.; Shaw, M. W. and Schull, W. J., ed., Public Health Service Publication No. 1163, Washington D.C., 45-60.

SCHULL, W. J. and ITO, P. K. (1969). A note on estimation of the ABO gene frequencies and the coefficient of inbreeding. Amer. J. Hum. Genet., 21, 168-170.

SMITH, C. A. B. (1957). Counting methods in genetical statistics. Ann. Hum. Genet., Lond., 21, 254-276.

SMITH, C. A. B. (1967). Notes on gene frequency estimation with multiple alleles. Ann. Hum. Genet., Lond., 31, 99-107.

STANTON, R. G. (1960). Genetic correlations with multiple alleles. Biometrics, 16, 235-244.

STERN, C. (1943). The Hardy-Weinberg Law. Science, 97, 137-138.

STERN, C. (1960). Principles of Human Genetics, Freeman and Co., San Francisco.

STEVENS, W. L. (1938). Estimation of blood group gene frequencies. Ann. Eugen., Lond., 8, 362-375.

STEVENS, W. L. (1950). Statistical Analysis of the ABO blood groups. Hum. Biol., 22, 191-217

STOCKS, P. (1952). Recent statistics of multiple births in England and Wales. Acta Genet. Med. et Gamell., 1, 8-13.

SUKHATME, P. V. (1942). On Bernstein's Improved method of estimating blood group gene frequencies. Sankhyā, 6, 85-92.

TAYLOR, G. L. and PRIOR, A. M. (1938). Blood groups in England. I. Examination of family and unrelated materials. Ann. Eugen., Lond., 8, 343-355.

WAHLUND, S. (1928). Cited in YASUDA, N. (1966).

WATERHOUSE, J. A. H. (1950). Twinning in twin pedigrees. Brit J. Soc. Med., 4, 197-216.

WEINBERG, W. (1901). Cited in DAS, S. R. (1953, 1955, 1956).

WEINBERG, W. (1909). Cited in LI, C. C. (1968).

WEINBERG, W. (1909). Cited in WATERHOUSE, J. A. H. (1950).

WEINBERG, W. (1910). Cited in LI, C. C. (1968).

WENTWORTH, E. N. and REMICK, B. L. (1916). Some breeding properties of the generalized Mendelian population. Genetics, 1, 608-616.

WHITE, C. and WYSHAK, G. (1964). Inheritance of human dizygotic twinning. New England Journ. Med., 271, 1003-1005.

WIENER, A. S.; LEDERER, M. and POLAYES, S. H. (1929). Studies in isohemagglutination. I. Theoretical considerations. J. Immun., Balt., 16, 469-482.

WIENER, A. S. (1954). Serology, genetics and nomenclature of the MNS types. Acta Genet. Med. et Gamell., 3, 314-321.

WRIGHT, S. (1918). On the nature of size factors. Genetics, 3, 367-374.

WRIGHT, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. Proc. Nat. Acad. Sci., 6, 320-332.

WRIGHT, S. (1921). Systems of Mating. Genetics, 6, 111-178.

WRIGHT, S. (1949). Adaptation and selection. In : Genetics, Paleontology and Evolution. Princeton Univ. Press, 365-389.

YASUDA, N. (1966). The genetical structure of northeastern Brazil. Ph.D. dissertation. University of Hawaii.

YASUDA, N. (1966). A singularity at the ABO blood group locus. Third Int. Cong. of Hum. Genet., Chicago, Abstracts of contributed papers. Abstract No. 353.

YASUDA, N. (1968). An extension of Wahlund's principle to evaluate mating type frequencies. Amer. J. Hum. Genet., 20, 1-23.

YASUDA, N. and KIMURA, M. (1968). A gene counting method of maximum likelihood for estimating gene frequencies in ABO and ABO like systems. Ann. Hum. Genet., Lond., 31, 409-420.

YEE, S. and MORTON, N. E. (1970). Letter to the editor. Amer. J. Hum. Genet., 22, 112-113.

YULE, G. U. (1924). Mathematical theory of evolution based on conclusions of Dr. J. C. Willis. Phil. Trans. Roy. Soc., B, 213, 21.

WARD, R. H. and SING, C. F. (1970). A consideration of the power of the chi-square test to detect inbreeding effects in natural populations. To appear in Amer. J. Hum. Genet.